On the Hasse Principle for Systems of Forms

MATTHEW JOSEPH NORTHEY

On the Hasse Principle for Systems of Forms

Matthew J. Northey

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
Durham University
United Kingdom

March 2022

On the Hasse Principle for Systems of Forms

Matthew J. Northey

Submitted for the degree of Doctor of Philosophy

March 2022

Abstract: We prove the Hasse principle for a smooth projective variety $X \subset \mathbb{P}^{n-1}_{\mathbb{Q}}$ defined by a smooth system of two cubic polynomials in $n \geq 39$ variables. The main tool here is the development of a version of Kloosterman refinement for a smooth system of equations defined over $\mathbb{Q}$.

Declaration

The work in this thesis is based on research carried out in the Department of Mathematical Sciences at Durham University. No part of this thesis has been submitted elsewhere for any degree or qualification.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Pankaj Vishe for his continued support and advice throughout my thesis. Pankaj has been particularly understanding with regards to my health difficulties, and without his support I probably would not have been able to complete this thesis.

I would also like that the maths department for the various ways in which they have accommodated me and made me feel welcome, particularly Prof. John Parker and the maths office staff. It has been a privilege for me to be able to complete my PhD at Durham.

There are also many others both within the department and outside of it who have encouraged me over the years, whom I would like to thank. There are too many people to name everyone, but I would particularly like to thank Dr John Blackman, who's friendship has been invaluable to me throughout my time at Durham.

Finally, I would like to thank my family for everything that they have done over the years, and the support they have given me.

Contents

Chapter 1

An Introduction to the Circle Method

This thesis seeks to improve upon the current cutting edge version of the circle method for systems of forms through the introduction of *averaged van der Corput Differencing* and *Kloosterman refinement* into this setting. In consideration for readers who are not familiar with the circle method, we will firstly introduce the type of problem that one applies this technique to in order to motivate its conception. This is what will be covered in this chapter. In Chapter 2, we will survey several results of significance and allude to the key ideas that underpin them. Following this, in Chapter 3, we will discuss the key results that will arise from this thesis and introduce a few important definitions. In Chapter 4, we will cover several important auxiliary results that will be needed in later chapters of this thesis.

After discussing the auxiliary results, we will begin the proof of the results discussed in Chapter 3 in earnest: In Chapter 5, we will rigorously set up the circle method for our particular context before delving into Weyl and van der Corput differencing in Chapters 6 and 7, respectively. The crux of the problem will lie in finding a good upper bound for the *minor arcs* which appear in Chapter 5. We will use Weyl differencing to directly find a non-trivial upper bound for the cubic exponential sums that appear in the minor arcs, whilst van der Corput differencing will be used to

bound these cubic sums in terms of quadratic sums. From here, our attention will move to finding a non-trivial upper bound for the quadratic exponential sums that appear due to van der Corput differencing. This will in turn give us a bound for the cubic exponential sums found in Chapter 5; this will be the topic of Chapters 8-10. These bounds will be attained via Poisson summation and careful counting arguments.

By combining the results found in the first ten chapters, it will be possible to find several different non-trivial upper bounds for the minor arcs defined in Chapter 5, and so our next task will be to explicitly state these bounds and apply them optimally. Unfortunately, these bounds will be very complicated, and this will make it difficult to work with them "by hand". We will therefore introduce the theory for an algorithm which will enable us to use a computer to work with these bounds in Chapter 11 before manipulating them into a form which is compatible with the algorithm from Chapter 12. This will then enable us to verify the upper bound that we need for the minor arcs via an actualisation of this algorithm built using software such as Python or Mathematica. The code for a suitable algorithm using Mathematica can be found in the appendix.

Finally, we will use standard procedures to find an asymptotic formula for the *major arcs* defined in Chapter 5. This will enable us to prove the main theorems in Chapter 3. The treatment of the major arcs will be the topic of Chapter 13.

1.1 Existence of Rational Solutions

The problem which will serve our motivation for developing the circle method is one of the oldest mathematical problems: If we have some polynomial in n variables, $f(\underline{x})$, with integer coefficients, is there a way to determine whether or not the equation $f(\underline{x}) = 0$ has a rational solution $\underline{x} \in \mathbb{Q}^n$?

In the case when $n = 1$ and $\deg(f) < 5$, this is a very easy problem to solve since we have an explicit formula for the complex solutions of f, and so in principle, one can

just check each of these solutions by hand to determine whether or not one of them is rational. However, when $\deg(f) \geq 5$, it was proven by Galois that no general formula exists, so this method cannot be used.

Worse still, when $n > 1$ we hit a more fundamental problem with this approach: In this case, there are infinitely many real/complex solutions of f, and so there is no way to check them all by hand in a finite amount of time. We therefore need a more sophisticated approach in order to make any real progress towards determining whether or not f has a rational solution for a general polynomial in n variables of degree d.

In this thesis (and when using the circle method more generally), we will restrict ourselves slightly and focus on *forms* in n variables of degree d. A form is a polynomial with integer coefficients, comprised entirely of monomials which have the same degree. For example $F(x, y, z) = xyz+2x^2y-z^3$ is a form of degree 3 because each monomial is cubic. Naturally for any form in n variables of degree d, $F(\underline{x})$, we have $F(\underline{0}) = 0$, so we will be searching for non-trivial solutions.

The starting point of the circle method is to give up on trying to find an explicit rational solution for F and instead just try to prove that a rational solution exists. We will actually try to prove existence of an integer solution: In particular, if we define the counting function

$$N_F(P) := \#\{\underline{x} \in \mathbb{Z}^n \ : \ |\underline{x}| \leq P, \, F(\underline{x}) = 0\}, \tag{1.1}$$

then F has a non-trivial integer solution if and only if there is some $P > 1$ such that $N_F(P) > 1$. In fact, if $n > d$, we expect that if $N_F(P)$ is non-zero (and as long as F is sufficiently "nice"), then it should be of size $O(P^{n-d})$ provided that P is chosen to be sufficiently large. Hence, we will aim to prove the asymptotic formula

$$N_F(P) = c_F P^{n-d} + O(P^{n-d-\delta}) \tag{1.2}$$

for some constant c_F, and some $\delta > 0$. Working with the definition of $N_F(P)$ directly

is not feasible, so we will find a different way to write $N_F(P)$. Indeed,

$$N_F(P) = \sum_{|\underline{x}| \leq P} \delta_F \tag{1.3}$$

where

$$\delta_F := \begin{cases} 1 \text{ if } F(\underline{x}) = 0, \\ 0 \text{ else,} \end{cases}$$

and it is well known that

$$\delta_F = \int_0^1 e(\alpha F(\underline{x}))\, d\alpha,$$

where $e(y) := e^{2\pi i y}$. Hence we may rewrite (1.3) as follows:

$$N_F(P) = \int_0^1 \sum_{|\underline{x}| \leq P} e(\alpha F(\underline{x}))\, d\alpha. \tag{1.4}$$

Note that we have managed to write $N_F(P)$ in terms of a contour integral about the unit circle; this is where *the circle method* gets its name from. The circle method was originally developed by Hardy and Littlewood in a series of papers on Waring's problem in the 1920's. Its original formulation used objects from complex analysis instead of exponential sums, but Vinogradov quickly noticed that several technical complexities could be removed if one worked with the latter.

The fact that there is an exponential sum within the integral is useful for us because they are reasonably well understood objects, and there are several techniques from analytic number theory which can be used to bound the size of such sums. In particular, we can be hopeful that there will be a lot of cancellation occurring in these sums (for most α) due to the cyclic nature of $e(y)$. Therefore, if we can determine which α make $S(\alpha) := \sum e(\alpha F(\underline{x}))$ relatively "small" (due to cancellation from terms in the sum), then we can expect those α to make up the error term in (1.2).

In the next section, we will discuss a heuristic to describe which α we expect to contribute to the main term and which α we expect to contribute to the error term of (1.2).

1.2 Square-Root Cancellations

The intention of this section is to help the reader understand why we expect certain values of α to make $|S(\alpha)|$ large, and we why expect the rest to make $|S(\alpha)|$ relatively small. This is not intended to be mathematically rigorous. For the moment, we will consider when α is rational, say $\alpha = a/q$. Then, provided that q is not too large (and assuming that F is "nice" in some way), we expect the following bound to be true:

$$|S(a/q)| := \Big| \sum_{|\underline{x}| \leq P} e_q(aF(x)) \Big| \ll P^n q^{-n/2}, \tag{1.5}$$

where $e_q(y) := e^{2\pi i y/q}$ and "$\ll$" means "less than some constant multiple of". To see some intuition as to why we can expect this, we will assume that F is homogeneous for now, and let $\underline{x} = q\underline{u} + \underline{v}$. Then

$$\begin{aligned} |S(a/q)| = \Big| \sum_{|\underline{x}| \leq P} e_q(aF(x)) \Big| &\approx \Big| \sum_{|\underline{u}| \leq P/q} \sum_{\underline{v} \bmod q} e_q(aF(\underline{v})) \Big| \\ &= P^n q^{-n} \Big| \sum_{\underline{v} \bmod q} e_q(aF(\underline{v})) \Big|. \end{aligned} \tag{1.6}$$

Now, let $S_{q,c} := \{\underline{v} \in (\mathbb{Z}/q\mathbb{Z})^n \ : \ F(\underline{v}) \equiv c \mod q\}$, and note that if $\#S_{q,c_i} = \#S_{q,c_j}$ for every $c_i, c_j \in \mathbb{Z}/q\mathbb{Z}$, then we would have $\sum_{\underline{v} \bmod q} e_q(aF(\underline{v})) = 0$ since we would just be summing over the q-th roots of unity q^{n-1} times. Naturally, we do not have $\#S_{q,c_i} = \#S_{q,c_j}$ in general, but it is reasonable to expect that these sets do not differ from each other very much since $F(\underline{x})$ should "hit" each value modulo q roughly the same number of times. If we indeed had $\#S_{q,c_i} \approx \#S_{q,c_j}$, this would lead to a non-trivial upper bound for $\sum e_q(aF(\underline{v})) = 0$.

Due to this cancellation from summing over the q-th roots of unity, it turns out that we can expect

$$\Big| \sum_{\underline{v} \bmod q} e_q(aF(\underline{v})) \Big| \ll q^{n/2},$$

which is significantly better than the non-trivial bound of q^n (provided that q is not small). Combining this with (1.6) gives us the expected bound $|S(a/q)| \ll P^n q^{-n/2}$.

Note 1.1. Throughout this thesis, we will use "$x \ll y$" to mean "x is less than some constant multiple of y", "$x \gg y$" to mean "x is greater than some constant multiple of y", and $x \asymp y$ to mean $y \ll x \ll y$.

1.2.1 Kloosterman refinement

Before moving on to discuss how the heuristic of square-root cancellations can be used to predict which α in (1.4) will contribute to the main term of (1.2) (and which α will contribute to the error term), we will briefly introduce an improved version of square-root cancellations known as Kloosterman refinement. Kloosterman's revolutionary idea is actually rather simple: Instead of trying to bound $|S(a/q)|$, he instead considered bounding

$$\hat{S}(q) := \sum_{\substack{a \bmod q \\ (a,q)=1}} \sum_{|\underline{x}| \leq P} e_q(aF(x)) = \sum_{\substack{a \bmod q \\ (a,q)=1}} S(a/q).$$

One can use the triangle inequality and (1.6) to get

$$|\hat{S}(q)| \leq \sum_{\substack{a \bmod q \\ (a,q)=1}} |S(a/q)| \ll P^n q^{1-n/2},$$

but this is quite wasteful since we are essentially just summing over the a sum trivially. Instead, Kloosterman realised that one can use averaging arguments to extend the idea of square-root cancellations to sums like $\hat{S}(q)$. In the context of his work [19], this enabled him to save an extra factor of $q^{1/2}$ which ultimately led him to the bound

$$|\hat{S}(q)| \leq \sum_{\substack{a \bmod q \\ (a,q)=1}} |S(a/q)| \ll P^n q^{1/2-n/2}.$$

It is difficult to explain the significance of this without seeing how the circle method operates in practice. However, we note that our ultimate goal is to derive the asymptotic formula, (1.2), for $N_F(P)$, and so it will naturally be helpful to have good upper bounds for sums relating to (1.4). In the context of this thesis, Kloosterman refinement will help us to improve the bounds related to the α's contributing to the

error term of (1.2).

1.3 Major and Minor Arcs

We now turn back to the problem of predicting which α in (1.4) will contribute to the main term of (1.2). The heuristic (1.5) tells us that we only expect $|S(a/q)|$ to be close to its trivial bound P^n when q is small. Therefore, by continuity of $S(\alpha)$, we only expect $|S(\alpha)|$ to be close to its trivial bound if $\alpha \in [0,1]$ is "close" to a rational number of low denominator.

This motivates the following decomposition of $[0,1]$: We will define

$$\mathfrak{M}(\Delta) := \bigcup_{\substack{a,q \leq P^{\Delta} \\ (a,q)=1}} \left\{ z \in [0,1] \ : \ \left| \frac{a}{q} - z \right| < P^{-d+\Delta} \right\}$$

to be the *major arcs* of $N_F(P)$, and $\mathfrak{m}(\Delta) := [0,1] \backslash \mathfrak{M}(\Delta)$ to be the *minor arcs* of $N_P(F)$, where Δ is some small constant. Then by (1.4), we may write

$$N_F(P) = \int_{\mathfrak{M}(\Delta)} S(\alpha) d\alpha + \int_{\mathfrak{m}(\Delta)} S(\alpha) d\alpha. \tag{1.7}$$

Remark 1.2. *As a brief aside, if Δ is chosen to be sufficiently small, the intervals of $\mathfrak{M}(\Delta)$ will be disjoint from one another, provided that P is large enough. Indeed, if we let $a_1/q_1, a_2/q_2 \in \mathbb{Q}$ be distinct fractions such that $q_1, q_2 \leq P^{\Delta}$, then*

$$\left| \frac{a_1}{q_1} - \frac{a_2}{q_2} \right| \geq \frac{1}{q_1 q_2} \geq P^{-2\Delta}. \tag{1.8}$$

We also note that

$$\left\{ z \in [0,1] \ : \ \left| \frac{a_1}{q_1} - z \right| < P^{-d+\Delta} \right\} \cap \left\{ z \in [0,1] \ : \ \left| \frac{a_2}{q_2} - z \right| < P^{-d+\Delta} \right\} = \phi$$

if and only if

$$\left| \frac{a_1}{q_1} - \frac{a_2}{q_2} \right| \geq \frac{1}{2} P^{-d+\Delta}.$$

This is certainly achieved for sufficiently large P if $\Delta < d/3$ by (1.8), proving that individual major arcs are pairwise disjoint from one another for such a choice of Δ.

Based on the rough heuristic that led to (1.5), we hope that the major arcs will give a contribution of order P^{n-d} and that the minor arcs will give an error term due to the expected extra cancellation in the exponential sums when we are not near to a rational of low denominator.

A simple reason why we chose P^{-d} in the definition of the major arcs is that we are expecting a contribution of (roughly) P^n to come from the exponential sums in this case, and so we need a contribution of roughly $O(P^{-d})$ to come from the integral in order for the major arcs to match up with our expected asymptotic formula. By continuity, $S(\alpha)$ will not change much over such a short region of α, so we expect that the integral will contribute its measure on the major arcs. This will be (roughly) P^{-d} since Δ will be chosen to be quite small.

In general, the circle method is used to prove that a polynomial F has integer solutions provided that the number of variables n is sufficiently large (and that F is "nice" in some way). Having a condition like this on n is not unreasonable since having more variables grants extra degrees of freedom with which one can try to force a solution. When we speak about improving/refining the circle method, we therefore usually mean finding more sophisticated ways to set the circle method up and/or finding better ways to bound $S(\alpha)$ so that we can weaken the assumption on n. The limiting factor for n usually lies with the minor arcs, as there are robust methods to show that the major arcs are of size $c_F P^{n-d} + O(P^{n-d-\delta})$, even when n is reasonably small relative to d.

Now that we have covered the basic set up of the circle method, we will discuss exactly what conditions are needed on $F(\underline{x})$ in order to be able to use it effectively.

1.4 Assumptions on $F(\underline{x})$ and the Hasse Principle

In this section, we aim to describe the type of result one expects to get when applying the circle method to this type of problem. In general we will make the following assumptions on our form F:

1. Assume that $F(\underline{x})$ is non-degenerate. This means we demand that there is no way to map the form F in n variables to another form G in $n-k$ variables, $k>0$, via a projective transformation. This is a natural assumption to have since we are trying to prove the F has an integer solution provided that F is in sufficiently many variables. We will therefore run into issues if F is actually a form in fewer variables in disguise.

2. Assume that $F(\underline{x})$ is absolutely irreducible. That is, we demand that F is irreducible over $\mathbb{C}$. This is again natural to some extent since if F was reducible, then F is actually a product of two polynomials of lower degree. It is not too surprising that an assumption like this is required since the condition on n relies on the degree of F.

3. Assume that $F(\underline{x})$ is non-singular over $\mathbb{C}$. That is, the set

$$\text{Sing}(F) := \{\underline{x} \in \mathbb{P}^{n-1}_{\mathbb{C}} \ : \ F(\underline{x}) = 0, \ \nabla F(\underline{x}) = \underline{0}\}$$

   is empty. When this set is non-empty, our exponential sum bounds become worse. This assumption is not strictly needed in order for the circle method to work, but we will assume this in order to avoid extra technical complications.

These three conditions being true is what we meant when we required F to be "nice" in Sections 1.2 - 1.3. It should be noted here that we will work with $\underline{x} \in \mathbb{P}^{n-1}_{\mathbb{Q}}$ from now on. It is natural to work in a projective space as opposed to an affine space since we are working with forms instead of regular polynomials. For example, we must

avoid the solution $\underline{x} = \underline{0}$ since $\underline{0}$ is not well defined in projective space. This will require us to change our counting function when we begin to use the circle method rigorously, but we will avoid these technical details in this section. We will discuss the actual counting function that we need to use in Chapter 3.

We also see that these conditions are not particularly restrictive: The first two conditions are there to ensure that we are genuinely working with a form in n variables, of degree d as opposed to something in fewer variables/of lower degree in disguise. As for the non-singularity condition, this is a more significant restriction, but it is an assumption that we do not strictly need to make.

Our ultimate goal will be to verify the *Hasse principle*. The simplest formulation of the Hasse principle is the following:

The Hasse Principle. If F has a real solution, and F has a p-adic solution for every p-adic field, $\mathbb{Q}_p$ (p prime), then F has a rational solution.

Broadly speaking, the Hasse principle tells us that the only way for F to not have a rational solution is due to it not having a real solution, or due to $F(\underline{x}) \equiv 0 \mod p^k$ having no solutions for some prime p, $k \in \mathbb{N}$. However, it should be noted that the converse of this is trivially true, and so whilst verifying the Hasse principle is not quite as good as proving that F has a rational solution, it is not far off in some sense. In particular if the Hasse principle is true, then we have precluded all possible ways for which F could fail to have a rational solution except for the most trivial way (no rational or p-adic solutions).

We are now able to state the type of result that one typically hopes for when using the circle method in this context:

Principle 1.3. *Let F be a form of degree d, in n variables. Assume that F is non-degenerate, non-singular, and absolutely irreducible. Then the Hasse principle is true provided that $n \geq c(d)$, where c is some constant depending only on d. In particular, there is some $\delta > 0$ such that*

$$N_F(P) = c_F P^{n-d} + O(P^{n-d-\delta}).$$

Going forward, we will not usually explicitly state that F is non-degenerate and absolutely irreducible since this is implicitly assumed by the fact that we want to work with a form that is genuinely in n variables of degree d.

We have now covered the basic motivation and set up going into the circle method. In the next chapter, we will survey several key results that have been discovered over the past century and we will briefly describe how the authors of these papers achieved these results. This context is necessary to understand how the work in this thesis improves upon the current cutting edge techniques.

Chapter 2

Survey of Results

In this chapter, we will briefly survey several results which are relevant to the techniques used in this thesis. We will start by discussing results related to applying the circle method to a single form before moving onto results related to systems of forms. This will help the reader to see how the current cutting edge techniques for a single form are superior to the best techniques for systems of forms, and it will enable us to discuss how the work in this thesis aims to partially bridge this gap.

The survey will not be ordered by date; results will instead be grouped together by the techniques that were used in order to achieve these results. Most techniques that are mentioned here will also appear in the main body of this thesis.

2.1 Results of Significance for a Single Form

Regarding the topic of verifying the Hasse principle for (systems of) forms, there is one very general result that can be thought of as a benchmark of sorts, and that is Birch's Theorem, which comes from his landmark paper in 1961 [1]. In the case of a single form F of degree d in n variables, Birch managed to show that the Hasse principle is true provided that

$$n - \sigma > (d-1)2^d + 1,$$

where $\sigma_{\mathbb{C}} = \sigma := \dim\{\underline{x} \in \mathbb{P}^{n-1}_{\mathbb{C}} \ : \ F(\underline{x}) = 0, \nabla F(\underline{x}) = 0\}$ is the dimension of the singular locus of F.

To get this result, Birch used an application of Weyl differencing to bound the exponential sum $S(\alpha)$ by a related exponential sum which is in terms of a linear polynomial instead of a polynomial of degree d. The reason why he did this is that it is in general much easier to find non-trivial bounds for linear exponential sums, and this ultimately led to him finding a non-trivial bound for the minor arcs.

One of the largest advantages to Weyl differencing is that it is relatively easy to use in a high level on generality, but there is a more sophisticated differencing argument in the literature known as van der Corput differencing. In 2014 Browning and Prendiville [5] managed to use a combination of van der Corput differencing and Weyl differencing to verify that the Hasse principle is satisfied provided that

$$n - \sigma > (d - \frac{1}{2}\sqrt{d})2^d + 1.$$

This – to the author's knowledge – is the best known result for an arbitrary form of degree d, as long as $d \geq 5$. In the case where d is small however, there are several other results.

For example, in 2007, Heath-Brown observed that it was possible to improve the bounds when performing van der Corput differencing by using an averaging argument over the minor arcs integral [13]. He used this to show that cubic forms must have an integer solution, provided that $n - \sigma \geq 15$.

Even though van der Corput differencing can be used to get significantly better results than Weyl differencing, it is quite unlikely that we will be able to improve van der Corput differencing sufficiently to remove the 2^d term from our lower bound on n, even though heuristically one would hope that the Hasse principle is true for $n - \sigma \geq d + 1$. The main issue is that using Weyl/van der Corput differencing to bound our non-linear exponential sum by a linear one turns out to be very wasteful. However, some progress has been made in this area:

In 2009 Browning and Heath-Brown showed that it was possible to improve upon Birch's Theorem when $d = 4$ by using van der Corput differencing to bound the related quartic exponential sum by a cubic sum (instead of a linear sum), and then applying the Poisson summation formula. This enabled them to verify the Hasse principle for $n - \sigma > 41$, which is an improvement of 8 variables over Birch. Hanselmann then managed to build on this work and attain $n - \sigma > 40$ by taking advantage of averaging over the minor arcs integral [10] in a similar spirit to [13].

So far, there has not been a way to attain a non-trivial upper bound using the Poisson summation formula directly when $d \geq 4$; one must first use van der Corput differencing repeatedly in order to work with a cubic exponential sum. If it were possible to directly use the Poisson summation formula (without differencing), then it is quite likely that this would enable us to get a lower bound for n which grows polynomially in d as opposed to the exponentially growing lower bound that we currently have from Browning and Prendiville.

Using the Poisson summation formula also comes with one additional advantage over using differencing methods to work with linear exponential sums: We may potentially be able to take advantage Kloosterman refinement which – if we can use it – allows a significant improvement to our bound on n. Currently there is no known way to use Kloosterman refinement effectively when working with linear exponential sums.

In 1996, Heath-Brown developed a technique known as the Delta Method which allows one to use Kloosterman refinement, and then found a way to use the Poisson summation formula to capitalise on this. In the paper *A New Form of the circle method and its Application to Quadratic Forms* [12], Heath-Brown used Kloosterman refinement to verify the Hasse principle for $n \geq 3$, which is the best possible result for a single quadratic form. He also managed to show that Cubic forms in 10 variables always have a rational solution [11] by attaining Kloosterman refinement via a different path than the Delta Method, which is again the best possible result of this type. Hooley then built on Heath-Brown's work to verify the Hasse principle

for a cubic form when $n \geq 9$ [16].

More recently, Vishe and Marmon [21] discovered a way to use the Delta Method whilst also using van der Corput differencing. They managed to combine the Delta method, van der Corput differencing (with integral averaging), and Poisson summation to verify the Hasse principle for a single quartic provided that $n-\sigma \geq 31$, which is an improvement of 10 variables over Hanselmann. In principle, their approach can be generalised to higher degree, and so this is currently the cutting edge version of the circle method for a single form of degree $d \geq 4$. We will now turn to results pertaining to systems of forms.

2.2 Results of Significance for Systems of Forms

In the case of systems of forms, much less is known. Birch's Theorem [1] can be thought of as a benchmark: In this higher level of generality, Birch managed to show that for a systems of R forms, $F_1, \cdots, F_R$ of degree d, the Hasse principle is true provided that

$$n - \sigma' > (d-1)R(R+1)2^{d-1} + 1,$$

where

$$\sigma' := \{\underline{x} \in \mathbb{P}^{n-1}_{\mathbb{C}} \ : \ \text{Rank}(\nabla F_1(\underline{x}), \cdots, \nabla F_R(\underline{x})) < R\}.$$

This set can be thought of as a more primitive version of a singular locus. It has taken over 50 years for anybody to find a result in a similar level of generality to this. This is due to several of the most powerful techniques used in the case of one form not having a known higher dimensional analogue (such as the Delta Method). There are also many extra complications which crop up even with the techniques that can be used. The first result of comparable generality to Birch's theorem that the author is aware of is a result of Lee: In 2011, Lee published a paper which proves a direct analogue to Birch's Theorem in the context of function fields [20].

In 2014, Heath-Brown and Browning also managed to generalise Birch's Theorem

to work with systems of forms of differing degree [4]. However, similarly to Lee, this result can be thought of as extending the number of contexts where Birch's Theorem is applicable, as opposed to a refinement of the techniques themselves.

To the author's knowledge, it was not until 2015 that somebody managed to directly improve Birch's theorem with a similar level of generality to Birch. In 2015-2017 Myerson found a way to use Weyl differencing more effectively when $R > 1$, and released a series of three papers which verify the Hasse principle provided that $n - \sigma' > 8R$ if $d = 2$ [26], $n - \sigma' > 25R$ if $d = 3$ [25], and

$$n > dR2^d + R$$

when $d \geq 4$ [24]. When $d \geq 4$ the system must also be "generic". This is a significant improvement over Birch's Theorem provided that R is not too small.

To the author's knowledge, van der Corput differencing has not been applied when $R > 1$, and Poisson summation has only been applied for special cases when R and d are both small. In particular, in 2015, Heath-Brown, Browning, and Dietmann used Poisson summation to show that a system of one cubic and one quadric form has rational solutions provided that their intersection is non-singular, and $n \geq 29$ [2].

In 2015, Munshi introduced a version of the Delta method which allows one to use Kloosterman refinement in the case of two quadrics [23]. He combined this with Poisson summation to verify the Hasse principle when $n \geq 11$, provided that their intersection is non-singular. Unfortunately, the techniques used are difficult to generalise effectively outside of the case of two quadrics.

Besides the result of Munshi, Vishe discovered a path to Kloosterman refinement when $R > 1$ in the function fields setting (forms over $\mathbb{F}_q(t)$) in 2019 [28]. Here he managed to verify the Hasse principle provided that the intersection of the forms is non-singular, and $n \geq 9$. So far, the technique used to introduce Kloosterman refinement in this context has not been extended to forms over $\mathbb{Q}$, but could be used more generally for forms over $\mathbb{F}_q(t)$.

We therefore see that many of the most powerful techniques used when $R = 1$ have not been extended to when $R > 1$, particularly when we are considering forms over $\mathbb{Q}$. In particular, both van der Corput Differencing (with averaging) and Kloosterman refinement (outside of a specific case) have not been used, and this is due to there being many extra difficulties which arise due to working with a system of forms instead a single form. In this thesis, we will aim to introduce a path to use Averaged van der Corput Differencing and Kloosterman refinement which will work in almost any context when $R > 1$. We will do this by studying the intersection of two cubic forms.

2.3 Results of Significance for Diagonal Forms

In this section, we will briefly touch on a special subset of the problem of verifying the Hasse principle for (systems of) forms, namely the case where our forms are *diagonal*. By this, we mean that our forms are of the following type:

$$F(\underline{x}) = c_1 x_1^d + c_2 x_2^d + \cdots + c_n x_n^d,$$

where $c_i \in \mathbb{Z}\backslash\{0\}$, $i \in \{1, \cdots, n\}$. In this case, the exponential sum $\sum_{\underline{x}} e(\alpha F(\underline{x}))$ becomes separable; in other words

$$\sum_{\underline{x}} e(\alpha F(\underline{x})) = \prod_{i=1}^{n} \sum_{x_i} e(\alpha c_i x_i^d).$$

The sums on the right are – in principle – far simpler objects to work with, and so in this special case, we can expect significantly stronger results than in the case where we are working with an arbitrary form of degree d.

For example, in 2016, Brüdern and Wooley showed that the Hasse principle is true for a system of diagonal cubic forms in R variables provided that $n > 6R$ [7], and provided that the coefficient matrix associated to the system has no singular $R \times R$ minor. This gives a modest improvement of two variables over Hooley's result in the general (not necessarily diagonal) case when $R = 1$, and a much more significant

improvement of thirty-six variables over Birch when $R = 2$. This result is particularly impressive as it achieves the theoretical limit of the circle method for such a system of forms.

Similarly, for non-singular systems of quartic diagonal forms, Brüdern and Wooley also verified the Hasse principle provided that $n \geq 22$ in the case of two quartics [6], and $n \geq 32$ in the case of three quartics [8]. In the case of three quartics, an additional technical condition is also required for the result to hold.

In the next chapter, we will state the main results that will arise from this thesis, as well as introduce a few necessary concepts. We will also lightly touch on Kloosterman refinement and allude to how we will capitalise on it in the context of a system of forms, but we will not go into detail until Chapters 5 and 7.

Chapter 3

Statement of Results

Let X denote a complete intersection variety in $\mathbb{P}^{n-1}_{\mathbb{Q}}$. Namely, X corresponds to the zero locus of a smooth system of R polynomials of degree d defined over $\mathbb{Q}$. Let

$$\sigma = \dim \mathrm{Sing}(X),$$

where

$$\mathrm{Sing}(X) := \{\underline{x} \in \mathbb{P}^{n-1}_{\mathbb{C}} \ : \ F_1(\underline{x}) = \cdots = F_R(\underline{x}) = 0, \ \mathrm{Rank}(\nabla F_1(\underline{x}), \cdots, \nabla F_R(\underline{x})) < R\} \tag{3.1}$$

denotes the singular locus of the variety X.

Furthermore, we define $\underline{x}_0$ to be a *smooth point* of X if

$$F_1(\underline{x}) = \cdots = F_R(\underline{x}) = 0, \quad \mathrm{Rank}(\nabla F_1(\underline{x}), \cdots, \nabla F_R(\underline{x})) = R. \tag{3.2}$$

The main purpose of this work is to provide a route to Kloosterman refinement for a system of forms over $\mathbb{Q}$ in the settings where the Poisson summation does not work directly. In particular, it should improve upon the current methods as long as X is not given by the two quadrics or an intersection a cubic and a quadric. We now define the setting in this thesis. Let $F(\underline{x}), G(\underline{x}) \in \mathbb{Z}[x_1, ..., x_n]$ be two homogeneous cubic forms in n variables and with integer coefficients, and let X denote the smooth projective variety defined by their simultaneous zero locus. In this case, the result

by Birch $n - \sigma \geq 50$ is yet to be improved. In this thesis, we will use Kloosterman refinement and a 2-dimensional version of averaged van der Corput differencing to improve upon Birch in the non-singular case.

In particular, we aim to prove the following result:

Theorem 3.1. *Let $X_{F,G} \subset \mathbb{P}^{n-1}_{\mathbb{Q}}$ be a complete intersection variety defined by a system of two cubic forms, and $X_F, X_G \subset \mathbb{P}^{n-1}_{\mathbb{Q}}$ be the varieties defined by F and G respectively. Let $\sigma(F,G)$, $\sigma(F)$, and $\sigma(G)$ be the respective dimensions of the singular loci of these varieties, and assume that*

$$m_\infty(F,G) := \max\{\sigma(F,G), \sigma(F), \sigma(G)\} = -1. \tag{3.3}$$

Then, the Hasse principle is true provided that $n \geq 39$.

This theorem can be generalised to $m_\infty(F,G) \geq -1$, but we will primarily focus on the non-singular case for simplicity in this thesis. To the best of the author's knowledge, this is the first known improvement of the Birch's result in this case. Let us briefly give an outline of the main idea of the proof. Our main tool in proving Theorem 3.1 is going to be presented by our main counting result in Theorem 3.2 below.

From now on, we will assume that X is a complete intersection of two cubics as before further satisfying:

$$X(\mathbb{A}_{\mathbb{Q}}) \neq \emptyset, \tag{3.4}$$

where

$$X(\mathbb{A}_{\mathbb{Q}}) := X(\mathbb{R}) \times \prod_p X(\mathbb{Q}_p).$$

This is saying that we are assuming that X has a real solution, and a solution in every p-adic field. Given a smooth weight function $\omega \in \mathrm{C}_c^\infty(\mathbb{R}^n)$, and a large parameter $1 \leq P$, we define the following smooth counting function:

$$N(P) := N_\omega(P) := \sum_{\substack{\underline{x} \in \mathbb{Z}^n, \\ F(\underline{x}) = G(\underline{x}) = 0}} \omega(\underline{x}/P).$$

Our main tool in proving Theorem 3.1 is the asymptotic formula for $N(P)$ obtained in Theorem 3.2. Before stating it, let us define the weight function ω in the following way. We will choose ω to be a smooth weight function, centred at a non-singular point $\underline{x}_0 \in X(\mathbb{R})$ with the additional property that its support is a "small" region about $\underline{x}_0$. Upon recalling (3.2), it is easy to see that the existence of such a point is guaranteed by our earlier assumptions that $X(\mathbb{R}) \neq \emptyset$ and that X is non-singular over $\mathbb{C}$. In particular any point $\underline{x}_0 \in X(\mathbb{R})$ must have $\mathrm{Rank}(\nabla F(\underline{x}_0), \nabla G(\underline{x}_0)) = 2$, otherwise $\mathrm{Sing}_{\mathbb{C}}(X) \neq \emptyset$ by definition. For convenience, set

$$\omega_P(\underline{x}) := \omega(\underline{x}/P).$$

There are two reasons why we choose the weight function in this way: Firstly, as alluded to in the Chapter 1, we need our counting function to avoid counting the origin since we are working in projective space. We achieve this for every P provided that the support of ω is sufficiently small, since $\mathrm{Supp}(\omega_P)$ will be a P-scaled version of $\mathrm{Supp}(\omega)$, centred at $P\underline{x}_0$ instead of $\underline{x}_0$. Demanding that ω be analytic will also make certain integrals easier to compute in future chapters.

Using homogeneity of F and G, we may further assume that $|\underline{x}_0| < 1$. This condition is superficial, and only assumed to make the implied constants appearing in our argument simpler. Let

$$\gamma(\underline{x}) := \begin{cases} \prod_j e^{-1/(1-|x_j|)^2} & \text{if } |\underline{x}| < 1, \\ 0 & \text{else,} \end{cases} \tag{3.5}$$

denote a non-negative smooth function supported in the hypercube $[-1,1]^n$. Given a parameter $0 < \rho < 1$ to be suitably decided later, we define

$$\omega(\underline{x}) := \gamma(\rho^{-1}(\underline{x} - \underline{x}_0)). \tag{3.6}$$

We are now set to state our main counting result, which directly implies Theorem 3.1.

Theorem 3.2. *Let $X \subset \mathbb{P}^{n-1}_{\mathbb{Q}}$ be a complete intersection variety defined by a system*

of two cubic forms F, G, such that $m_\infty(F,G) = -1$ (as defined in (3.3)). Then as long as $n \geq 39$ and $X_{\mathrm{ns}}(\mathbb{A}_{\mathbb{Q}}) \neq \emptyset$, there exist $C_X > 0$ and some $\rho_0 \in (0,1]$, such that for each $0 < \rho \leq \rho_0$, there exists $\delta_0 := \delta_0(\rho) > 0$ such that

$$N(P) = C_X P^{n-6} + O_{n,F,G,\rho}(P^{n-6-\delta_0}).$$

The circle method begins with by writing the counting function $N(P)$ as an integral of a suitable exponential sum:

$$N_\omega(P) = \sum_{\substack{\underline{x} \in \mathbb{Z}^n, \\ F(\underline{x}) = G(\underline{x}) = 0}} \omega(\underline{x}/P) = \int_0^1 \int_0^1 S(\alpha_1, \alpha_2) d\alpha_1 d\alpha_2,$$

where

$$S(\underline{\alpha}) := S(\alpha_1, \alpha_2) := \sum_{\underline{x} \in \mathbb{Z}^n} \omega(\underline{x}/P) e(\alpha_1 F(\underline{x}) + \alpha_2 G(\underline{x})), \tag{3.7}$$

denotes the corresponding exponential sum.

In the traditional circle method, the unit square $I := [0,1]^2$ is split into major arcs $\mathfrak{M}$ which consist of the points in I which are "close" to a rational point $\underline{a}/q$, where $\underline{a} = (a_1, a_2) \in \mathbb{Z}^2$ of "small" denominator q, and minor arcs $\mathfrak{m} = I \backslash \mathfrak{M}$ which consist of everything else. The limitation of the process usually occurs while bounding the integral

$$\int_{\mathfrak{m}} S(\underline{\alpha}) d\underline{\alpha}.$$

When $R = 1$, Kloosterman's revolutionary idea [19] was to use Farey fractions to partition $[0,1]$ to bound the minor arc contribution, usually called a Farey dissection. This idea essentially allows us – upon setting $\alpha := a/q + z$ and fixing the value of z – to consider averages of the corresponding exponential sum of the form

$$\sum_{\substack{a \bmod q \\ (a,q)=1}} S(a/q + z).$$

The extra average over a allows us to save an extra factor of size $O(q^{1/2})$, when q is sufficiently large and z relatively small (recall that we normally hope to save $O(q^{n/2})$ from $S(a/q+z)$ due to square root cancellations). We will carefully set up the circle

method and perform this Farey dissection of the minor arcs in Chapter 5.

When $R = 2$, it is quite difficult to find an analogue of the Farey fraction dissection which can be used to attain Kloosterman refinement, especially over $\mathbb{Q}$. In [28], Vishe managed find such an analogue in the function field setting, but so far it is not known how to use these ideas when working over $\mathbb{Q}$. The path to Kloosterman refinement in this thesis will not focus on innovations to the Farey dissection, and will instead focus on improving van der Corput differencing.

In the setting of that we will discuss (pair of two cubics), the Poisson summation formula cannot be applied directly. To be more precise, it is possible to apply Poisson summation, but the bound that it gives is trivial due to a certain integral bound behaving badly when the degrees of our forms become too large.

We therefore must use a differencing argument (such as van der Corput) to bound $|S(\alpha)|$ by a sum with polynomials of lower degree. To do this, one essentially starts by using Cauchy's inequality to bound

$$\left|\int_{\mathfrak{m}} S(\underline{\alpha})d\underline{\alpha}\right| \ll \left(\int_{\mathfrak{m}} |S(\underline{\alpha})|^2 d\underline{\alpha}\right)^{1/2}. \tag{3.8}$$

This leads us for a fixed integer q and a fixed small $\underline{z} \in I$ to consider the averages of the form

$$\int_{|\underline{z}|<q^{-1}Q^{-1/2}} \sum_{\substack{\underline{a} \bmod q \\ (\underline{a},q)=1}} |S(\underline{a}/q + \underline{z})|^2 d\underline{z}, \tag{3.9}$$

where Q is a suitable parameter to be fixed later. This parameter Q arises from using the two dimensional Dirichlet approximation theorem. We further develop a two dimensional version of averaged van der Corput differencing used by Hanselmann [10], and Marmon and Vishe [21] to estimate the averages of $|S(\underline{a}/q+\underline{z})|^2$ over $\underline{z}$. This leads us to considering quadratic exponential sums for a system of differenced quadratic forms

$$F_{\underline{h}}(\underline{x}) := \underline{h} \cdot \nabla F(\underline{x}),\ G_{\underline{h}}(\underline{x}) := \underline{h} \cdot \nabla G(\underline{x}). \tag{3.10}$$

The extra averaging over $\underline{a}$ in (3.9) leads us to a saving of the size $O(q)$ in the estimation of $\sum_a |S(\underline{a}/q + \underline{z})|^2$, and in the light of squaring technique used in (3.8),

it overall saves us a factor of size $O(q^{1/2})$ when q is square-free.

The methods developed here are versatile and can be readily adapted to deal with general complete intersections. While dealing with averages of squares of corresponding exponential sums next rationals of type $(a_1, ..., a_R)/q$, where q is square-free, we would be able to save a factor of size $O(q^{R/4})$ over the bounds coming from averaged van der Corput along with pointwise Poisson summation. To the best of the author's knowledge, this is the first known version of Kloosterman refinement which generalises this way over $\mathbb{Q}$. In the function field setting, this method could potentially be combined with the method of Vishe [28] to be able to save a factor of size $O(q^{(R-1)/4+1/2})$ instead.

Chapter 4

Background on a pair of quadrics

Exponential sums for a pair of quadrics will feature prominently in this work. Let $Q_1(\underline{x}), Q_2(\underline{x})$ be a pair of quadratic forms in n variables with integer coefficients and consider the variety defined by

$$V : Q_1(\underline{x}) = Q_2(\underline{x}) = 0,$$

$\underline{x} \in \overline{\mathbb{Q}}^n$. Let $\mathrm{Sing}_K(V)$ to be the (projective) singular locus of V over field K. When Q_1 and Q_2 intersect properly, namely, if V is of projective dimension $n-3$ then we can express the singular locus of V as follows:

$$\mathrm{Sing}_K(V) := \left\{ \underline{x} \in \mathbb{P}^{n-1}_K \;\middle|\; \underline{x} \in V,\ \mathrm{Rank} \begin{pmatrix} \nabla Q_1(\underline{x}) \\ \nabla Q_2(\underline{x}) \end{pmatrix} < 2 \right\}. \tag{4.1}$$

We say that the intersection variety of $Q_1(\underline{x})$ and $Q_2(\underline{x})$, V, is non-singular if $\dim \mathrm{Sing}_K(V) = -1$, and singular otherwise. It should be noted that (4.1) only truly encapsulates the set of singular points when Q_1, Q_2 have a *proper* intersection over K (that is, the polynomials $Q_1(\underline{x})$, $Q_2(\underline{x})$ share no common factor over K). However, $\mathrm{Sing}_K(V)$ is still a well defined set with a well defined dimension, even when Q_1 and Q_2 intersect improperly. In fact, $\dim \mathrm{Sing}_K(V)$ happens to be very large when Q_1 and Q_2 intersect improperly, and knowing this will be useful to us later when we are deriving our exponential sum bounds in Chapter 8. This will be the topic of the

first lemma of this section.

Lemma 4.1. *Let F, G either be a pair of quadrics which intersect improperly over a field K, and define*

$$\mathrm{Sing}_K(F) := \left\{ \underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \;\middle|\; F(\underline{x}) = 0,\ \nabla F(\underline{x}) = \underline{0} \right\},$$

$$m_K(F,G) := \max\{\dim \mathrm{Sing}_K(F), \dim \mathrm{Sing}_K(G), \dim \mathrm{Sing}_K(F,G)\}.$$

Then

$$m_K(F,G) = \begin{cases} n-1 & \text{if } F \equiv 0 \text{ or } G \equiv 0 \\ n-2 & \text{else.} \end{cases}$$

Proof. We trivially have $m_K = n-1$ if either $F \equiv 0$ or $G \equiv 0$. If this is not the case, then there are two possible ways for F, G to intersect improperly: $F \equiv \lambda G \not\equiv 0$ for some $\lambda \in K$, or $F = L_1 L_2$, $G = L_1 L_3$ for some lines L_i. If $F \equiv \lambda G$, then

$$\mathrm{Sing}_K(F,G) = \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ :\ F(\underline{x}) = 0\}$$

by (4.1), and this clearly has dimension $n-2$ since F is not the zero polynomial. Alternatively, if $F = L_1 L_2$, $G = L_1 L_3$, then let $L_i(\underline{x}) := \underline{c}_i \cdot \underline{x}$. It is easy to check that

$$\nabla F(\underline{x}) = L_2(\underline{x})\underline{c}_1 + L_1(\underline{x})\underline{c}_2, \quad \nabla G(\underline{x}) = L_3(\underline{x})\underline{c}_1 + L_1(\underline{x})\underline{c}_3.$$

Hence, when $L_1(\underline{x}) = 0$, there exists some $\lambda_{\underline{x}}$ such that $\nabla F(\underline{x}) = \lambda_{\underline{x}} \nabla G(\underline{x})$ and so

$$\mathrm{Rank}_K \begin{pmatrix} \nabla F(\underline{x}) \\ \nabla G(\underline{x}) \end{pmatrix} < 2.$$

Furthermore, when $L_1(\underline{x}) = 0$, $F(\underline{x}) = G(\underline{x}) = 0$. Hence

$$\{L_1(\underline{x}) = 0\} \subset \mathrm{Sing}_K(F,G).$$

But, $\dim_K\{L_1(\underline{x}) = 0\} = n-2$, and so $\dim \mathrm{Sing}_K(F,G) \geq n-2$, giving us $m_K(F,G) \geq n-2$. We also have that $m_K(F,G) \leq n-2$ since F, G are assumed to not be the zero polynomial. □

In Chapters 8 and 9, we will consider sums of the form

$$\sum_{\underline{a}}^{q\,*} \Big| \sum_{\underline{x}}^{q} e_q(a_1 F(\underline{x}) + a_2 G(\underline{x}) + \underline{m} \cdot \underline{x}) \Big|, \tag{4.2}$$

where the "star" in the first sum indicates that $(\underline{a}, q) = 1$, F, G are quadratic polynomials, and $\underline{m} \in \mathbb{Z}^n$ is some constant.

By absorbing the linear part of F and G into the constant $\underline{m}$ for a given $\underline{a}$ (and factorising out the constant part), we can view this object as an exponential sum constructed from a linear combination of two forms of the same degree. It is natural that at some points in the thesis, we will need to consider this linear combination as a single form in its own right. When we do this, we will aim to bound the size of its singular locus by the size of the singular locus of the intersection variety, $\mathrm{Sing}_K(V)$. We therefore now turn to generalising [14, Proposition 2.1]. The argument used there generalises directly, however here, we will use [21, Lemma 4.1]. This result will be used at later stages of this thesis as well, therefore we begin by reproducing it in our context:

Lemma 4.2. *Let Q_1, Q_2 be a pair of quadratic forms defining a complete intersection $X = V(Q_1, Q_2)$. Let Π be a collection of primes such that $\#\Pi = r \geq 0$ and define $\Pi_a := \{p \in \Pi \,|\, p > a\}$ for every $a \in \mathbb{N}$. Then there exists a constant $c' = c'(n)$ and a set of primitive linearly independent vectors*

$$\underline{e}_1, \cdots, \underline{e}_n \in \mathbb{Z}^n$$

satisfying the following property for any integer $0 \leq \eta \leq n-1$, any subset $\phi \neq I \subset \{1, 2\}$ and any $\upsilon \in \{\infty\} \cup \Pi_{2c'}$: The subspace $\Lambda_\eta \subset \mathbb{P}^{n-1}_{\mathbb{F}_\upsilon}$ spanned by the images of $\underline{e}_1, \cdots, \underline{e}_{n-\eta}$ is such that

$$\dim(X_I \cap \Lambda_\eta)_\upsilon = \max\{-1, \dim(X_I)_\upsilon - \eta\} \tag{4.3}$$

and

$$\dim \mathrm{Sing}((X_I \cap \Lambda_\eta)_\upsilon) = \max\{-1, \dim \mathrm{Sing}((X_I)_\upsilon) - \eta\}. \tag{4.4}$$

Here given $\emptyset \neq I \subset \{1,2\}$, let X_I denote the complete intersection variety defined by the forms $\{F_i : i \in I\}$. Moreover, the basis vectors $\underline{e}_i$ can be chosen so that

$$L/2 \leq |\underline{e}_i| \leq L \tag{4.5}$$

for every $i = 1, \cdots, n$ and

$$L^n \ll det(\underline{e}_1, \cdots, \underline{e}_n) \ll L^n \tag{4.6}$$

for some constant $L = O_n(r+1)$.

Proof. Note that the statement of this lemma is identical to that of [21, Lemma 4.1] except that in the latter there is an additional assumption that the closed subscheme $X_I \subset \mathbb{P}^{n-1}_{\mathbb{Z}}$ defined by $F_i = 0$ for all $i \in I$ satisfies

$$\dim(X_I)_v = n - 1 - |I|. \tag{4.7}$$

This is equivalent to the case when X_1 and X_2 intersect properly. Therefore, it is enough to consider different cases where we have an improper intersection. In each of these particular cases, somewhat softer argument works.

In the trivial case when $Q_1 = Q_2 = 0$, any basis $\underline{e}_1, ..., \underline{e}_n$ will work.

When $Q_2 = \lambda Q_1$, where $\lambda \in K$ and Q_1 a non zero quadratic form then we may apply [21, Lemma 4.1] only to the hypersurface X_1 to find a basis $\underline{e}_1, ..., \underline{e}_n$ which is chosen such that (4.3) and (4.4) hold for $I = \{1\}$. This choice will clearly work for all $I \subset \{1,2\}$.

In the remaining case when $Q_1 = L_1L_2, Q_2 = L_1L_3$, where $L_i = \underline{v}_i \cdot \underline{x}$ and L_2 is not a scalar multiple of L_3. In this case, it is easy to check that the singular locus of $X_1 \cap X_2$ to is the hyperplane $L_1 = 0$. Here, we may apply [21, Lemma 4.1] to the single variety defined by the cubic form $L_1L_2L_3 = 0$. The basis Λ that we get from this process will work here as well. □

Now, since Q_1 and Q_2 are quadratic forms, we may define M_1, M_2 to be their

respective associated coefficient matrices defined as follows: If

$$Q_i(\underline{x}) := \sum_{j=1}^{n}\sum_{k=j}^{n} b_{j,k}^{(i)} x_j x_k,$$

then

$$(M_i)_{j,k} := \begin{cases} b_{j,k}^{(i)} \text{ if } j = k \\ \frac{1}{2} b_{j,k}^{(i)} \text{ if } j < k \\ \frac{1}{2} b_{k,j}^{(i)} \text{ if } j > k. \end{cases} \tag{4.8}$$

We clearly have that $M_1, M_2 \in M_n(\mathbb{Z}/2)$ – the set of $n \times n$ matrices with coefficients of the form $a/2$, $a \in \mathbb{Z}$) – since $b_{k,j} \in \mathbb{Z}$. In this chapter, we will assume without loss of generality that $M_1, M_2 \in M_n(\mathbb{Z})$. This is because even if $M_1, M_2 \notin M_n(\mathbb{Z})$, we certainly have $2M_1, 2M_2 \in M_n(\mathbb{Z})$, and so we may work with $2Q_1$, $2Q_2$ and relabel instead. We are now ready to prove the following generalisation of [14, Proposition 2.1].

Proposition 4.3. *Let ν either denote a finite prime $\nu \gg_n 1$ or the infinite prime, let $\mathbb{F}_\nu$ either denote the corresponding finite field or $\mathbb{Q}$, and let*

$$m_\nu := \max\{\dim \text{ Sing}_{\mathbb{F}_\nu}(X_1), \dim \text{ Sing}_{\mathbb{F}_\nu}(X_2), \dim \text{ Sing}_{\mathbb{F}_\nu}(V)\}, \tag{4.9}$$

where X_i is the variety defined by $Q_i(\underline{x}) = 0$ and V is defined as above. Moreover for every $(a_1, a_2) \in \mathbb{F}_\nu^2 \backslash (0,0)$, the rank of the matrix associated to the quadratic form $a_1Q_1 + a_2Q_2$, $a_1M_1 + a_2M_2$, satisfies

$$\text{Rank}(a_1M_1 + a_2M_2) \geq n - m_\nu - 2. \tag{4.10}$$

Moreover, there exists a set of eigenvalues $\Gamma = \{\gamma_1, ..., \gamma_k\} \subset \overline{\mathbb{F}}_\nu$, such that as long as $a_1 \neq \lambda_i a_2$ for some $1 \leq i \leq k$,

$$\text{Rank}(a_1M_1 + a_2M_2) \geq n - m_\nu - 1.$$

Proof. Let M_1 and M_2 denote the integer matrices defining the forms Q_1 and Q_2 respectively. We firstly note that for $m_\nu = -1$, we recover (4.10) from [14, Proposition 2.1]. In this case, since M_1 and M_2 are invertible, if $\text{Rank}_\nu(a_1M_1 + a_2M_2) < n$,

then $a_1, a_2 \neq 0$. Therefore, the matrix $a_1M_1 + a_2M_2$ must be singular and so there is some $\underline{x} \in \overline{\mathbb{F}}_\nu$ such that

$$\begin{aligned} &(a_1M_1 + a_2M_2)\underline{x} = \underline{0} \\ \Leftrightarrow\ &a_2M_2\underline{x} = -a_1M_1\underline{x} \\ \Leftrightarrow\ &(M_1^{-1}M_2)\underline{x} = -a_1a_2^{-1}\underline{x}. \end{aligned}$$

Hence $-a_1a_2^{-1}$ must be an eigenvalue of $M_1^{-1}M_2$. There must be at most n such eigenvalues, and therefore Γ could be taken as the set of negatives of these eigenvalues.

If $m_\nu \neq -1$, we invoke Lemma 4.2. As long as $\nu \gg_n 1$, we obtain a basis $\underline{e}_1, ..., \underline{e}_n$ of $\mathbb{F}_\nu^n$ such that the system of quadrics Q_1', Q_2' corresponding to the restriction of Q_1 and Q_2 onto the subspace $\Lambda_{n-m_\nu-1}$ obeys (4.3) - (4.4). This clearly defines a system of non-singular quadratic forms defined over $n - m_\nu - 1$, whose complete intersection is non-singular over $\overline{\mathbb{F}}_\nu$ as well. Now let M_1' and M_2' denote the integer matrices defining the forms Q_1', and Q_2' respectively. The Lemma now follows from noticing that

$$\mathrm{Rank}(a_1M_1 + a_2M_2) \geq \mathrm{Rank}(a_1M_1' + a_2M_2'),$$

for any pair $(a_1, a_2) \in \mathbb{F}_\nu^2 \setminus (0, 0)$ and further using our analysis of the non-singular case above. □

Whilst, our main exponential sums bound will be found by using Poisson summation, this bound will only be effective when q is relatively large, so we will aim to supplement our Poisson bounds using Weyl differencing. However, when one performs Weyl differencing, the bound that is attained will use a 'Birch-type' singular locus instead of the more natural singular locus definition, (4.1).

Furthermore, when we are dealing with the major arcs, we will need to consider the singular locus of a linear combination of two cubic forms. In this next proposition, we aim to bound the dimension of these objects.

Proposition 4.4. *Let F, G be non-constant forms of any degree, K be a field, and*

let

$$\sigma_K(F) := \dim\{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ F(\underline{x}) = 0,\, \nabla F(\underline{x}) = \underline{0}\} \tag{4.11}$$

$$\sigma'_K(F,G) := \dim\{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ \mathrm{Rank} \begin{pmatrix} \nabla F(\underline{x}) \\ \nabla G(\underline{x}) \end{pmatrix} < 2\} \tag{4.12}$$

$$\sigma_K(F,G) := \dim \mathrm{Sing}_K(F,G). \tag{4.13}$$

Then, we have

$$\sigma_K(a_1F + a_2G) \leq \sigma'_K(F,G) \leq \sigma_K(F,G) + 1$$

for any $(a_1, a_2) \in K\backslash\{(0,0)\}$.

Proof. The proof of the first inequality is very simple

$$\begin{aligned} \underline{x} \in \mathrm{Sing}_K(a_1F + a_2G) &\implies \nabla(a_1F(\underline{x}) + a_2G(\underline{x})) = 0 \\ &\implies \mathrm{Rank}(\nabla F(\underline{x}), \nabla G(\underline{x})) < 2. \end{aligned}$$

Hence we automatically have $\sigma_K(a_1F + a_2G) \leq \sigma'_K(F,G)$. We now aim to show that $\sigma'_K(F,G) \leq \sigma_K(F,G) + 1$, noting that this is a generalisation of [4, Lemma 3.1] in the context of two forms. To prove this inequality, we will decompose $\mathrm{Sing}_K(F,G)$ and

$$\mathrm{Sing}'_K(F,G) := \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ \mathrm{Rank} \begin{pmatrix} \nabla F(\underline{x}) \\ \nabla G(\underline{x}) \end{pmatrix} < 2\}$$

into three sets each, and work with those instead. Since $\nabla F(\underline{x}) = \underline{0}$ implies that $F(\underline{x}) = 0$ (similar for G), we see that

$$\begin{aligned} \mathrm{Sing}_K(F,G) =& \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ G(\underline{x}) = 0,\, \nabla F(\underline{x}) = 0\} \\ &\cup \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ F(\underline{x}) = 0,\, \nabla G(\underline{x}) = 0\} \\ &\cup \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ F(\underline{x}) = G(\underline{x}) = 0, \exists \lambda \in K\backslash\{0\} \\ &\qquad\qquad \text{s.t. } \nabla F(\underline{x}) = \lambda \nabla G(\underline{x})\} \\ =:& S_1 \cup S_2 \cup S_3, \end{aligned}$$

where S_1, S_2, S_3 are defined in the obvious way. Similarly

$$\begin{aligned}\mathrm{Sing}'_K(F,G) =& \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ \nabla F(\underline{x}) = 0\} \\ & \cup \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ \nabla G(\underline{x}) = 0\} \\ & \cup \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ \exists \lambda \in K\backslash\{0\} \text{ s.t. } \nabla F(\underline{x}) = \lambda \nabla G(\underline{x})\} \\ =:& S'_1 \cup S'_2 \cup S'_3.\end{aligned}$$

Next, we note that

$$\sigma_K(F,G) = \max\{\dim S_1, \dim S_2, \dim S_3\}, \tag{4.14}$$

$$\sigma'_K(F,G) = \max\{\dim S'_1, \dim S'_2, \dim S'_3\}, \tag{4.15}$$

and so, if we can show that $\dim S'_i \leq \dim S_i + 1$ for every i, then we will be done. This is because $\dim S_i \leq \sigma_K(F,G)$ for every i by (4.14). We start with $i = 1$: Upon noting that $S_1 = S'_1 \cap \{G(\underline{x}) = 0\}$, we may use the affine dimension theorem to conclude that

$$\begin{aligned}\dim S'_1 &\leq \dim S_1 - \dim\{G(\underline{x}) = 0\} + n \\ &\leq \dim S_1 + 1,\end{aligned}$$

as required. We similarly get $\dim S'_2 \leq \dim S_2 + 1$ by the same argument. Finally, in the case of $i = 3$, we note that since F, G are forms, Euler's formula gives us

$$\nabla F(\underline{x}) = \lambda \nabla G(\underline{x}) \implies F(\underline{x}) = \lambda G(\underline{x}).$$

Therefore, if $F(\underline{x}) = 0$, then we automatically must have $G(\underline{x}) = 0$ since $\lambda \neq 0$. In particular this implies that

$$\begin{aligned}S_3 =& \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ F(\underline{x}) = G(\underline{x}) = 0, \exists \lambda \in K\backslash\{0\} \\ & \qquad\qquad \text{s.t. } \nabla F(\underline{x}) = \lambda \nabla G(\underline{x})\} \\ =& \{\underline{x} \in \mathbb{P}^{n-1}_{\overline{K}} \ : \ F(\underline{x}) = 0, \exists \lambda \in K\backslash\{0\} \text{ s.t. } \nabla F(\underline{x}) = \lambda \nabla G(\underline{x})\} \\ =& S'_3 \cap \{F(\underline{x}) = 0\}.\end{aligned}$$

Hence, by the affine dimension theorem, we have

$$\begin{aligned}\dim S_3' &\leq \dim S_3 - \dim\{F(\underline{x}) = 0\} + n \\ &\leq \dim S_3 + 1,\end{aligned}$$

Hence by (4.14) - (4.15), we have

$$\begin{aligned}\sigma_K'(F,G) &= \max\{\dim S_1',\ \dim S_2',\ \dim S_3'\} \\ &\leq \max\{\dim S_1,\ \dim S_2,\ \dim S_3\} + 1 \\ &= \sigma_K(F,G) + 1,\end{aligned}$$

as required.

□

Our main exponential sum bound will be in terms of the size of the null set

$$\mathrm{Null}_q(M) := \{\underline{x} \in (\mathbb{Z}/q\mathbb{Z})^n : M\underline{x} = \underline{0}\}, \tag{4.16}$$

for some matrix M. The following three Lemmas will be related to this set.

Lemma 4.5. *Let M_n be the set of $n \times n$ integer matrices. Then for every $u, v \in \mathbb{N}$, and every $M \in M_n(\mathbb{Z})$, we have*

$$\#\mathrm{Null}_{uv}(M) \leq \#\mathrm{Null}_u(M)\#\mathrm{Null}_v(M),$$

with equality if $(u, v) = 1$.

Proof. It is easy to prove that $\#\mathrm{Null}_q(M)$ is a multiplicative function, so we will not prove that

$$\#\mathrm{Null}_{uv}(M) = \#\mathrm{Null}_u(M)\#\mathrm{Null}_v(M) \tag{4.17}$$

when $(u, v) = 1$. We will be brief when showing the inequality, as this is a standard Hensel Lemma type of argument. If $\underline{x} \in \mathrm{Null}_{uv}(M)$, then we must have $\underline{x} \in \mathrm{Null}_u(M)$. Hence, if we write $\underline{x} := \underline{y} + u\underline{z}$, then $\underline{y}$ must be in $\mathrm{Null}_u(M)$.

Now, fix $\underline{y}$ and assume that there is some $\underline{z}_1, \underline{z}_2$ (not necessarily distinct) such that $\underline{y} + u\underline{z}_i \in \mathrm{Null}_{uv}(M)$. Then

$$M(\underline{y} + u\underline{z}_i) \equiv \underline{0} \mod uv,$$

and so

$$M(\underline{y} + u\underline{z}_2) - M(\underline{y} + u\underline{z}_1) = uM(\underline{z}_2 - \underline{z}_1) \equiv \underline{0} \mod uv.$$

Therefore, upon letting $\underline{z}_2 := \underline{z}_1 + \underline{z}'$ we must have

$$M\underline{z}' \equiv \underline{0} \mod v.$$

Hence, there can only be at most $\#\mathrm{Null}_v(M)$ possible values for $\underline{z}'$ and so there can only be at most $\#\mathrm{Null}_v(M)$ values for $\underline{z}$ such that $\underline{y} + u\underline{z} \in \mathrm{Null}_{uv}(M)$ for any given $\underline{y}$. We also have that $\underline{y}$ must be in $\mathrm{Null}_u(M)$. This gives us

$$\#\mathrm{Null}_{uv}(M) \leq \#\mathrm{Null}_u(M)\#\mathrm{Null}_v(M),$$

as required. □

In both Chapter 8 and 9, we will need to bound $\#\mathrm{Null}_p(M)$ for matrices of the form $M(\underline{a}) := a_1M_1 + a_2M_2$, where M_1 and M_2 are symmetric matrices associated to some quadratic forms $Q_1(\underline{x})$, $Q_2(\underline{x})$. In Proposition 4.3, we noted that for most values of $\underline{a}$, $\mathrm{Rank}_p(M(\underline{a})) \geq n - m_p - 1$, but there were potentially a few lines of $\underline{a}$'s where $\mathrm{Rank}_p(M(\underline{a})) = n - m_p - 2$. Naturally, a lower bound on the size of the rank of a matrix leads to an upper bound on the dimension of the nullspace of a matrix (due to the rank-nullity theorem), and so using $\mathrm{Rank}_p(M(\underline{a})) \geq n - m_p - 2$ in order to bound $\#\mathrm{Null}_p(M(\underline{a}))$ for every $\underline{a}$ would be wasteful. This will lead us to considering averages of $\#\mathrm{Null}_p(M(\underline{a}))$, where $\underline{a}$ is allowed to vary. This is the topic of the next lemma.

Lemma 4.6. *Let Q_1, Q_2 be quadratic forms in n variables, $q \in \mathbb{N}$, and*

$$d := \prod_{i=1}^{r} p_i$$

be squarefree (in other words, the p_i's are prime) such that $d \mid q$. Furthermore, let M_1, M_2 be integer matrices defining Q_1 and Q_2 respectively, and let $m_p = m_p(Q_1, Q_2)$ be as defined in (4.9) *for $K = \mathbb{F}_p$, p a prime. Then*

$$S(d,q) := \sideset{}{^*}\sum_{\underline{a} \bmod q} \#\mathrm{Null}_d(a_1M_1 + a_2M_2) \ll_n q^2 \prod_{i=1}^{r} p_i^{m_{p_i}+1}.$$

Proof. We firstly note that upon setting $\underline{a} = \underline{b} + d\underline{c}$,

$$\begin{aligned} S(d,q) &= \sideset{}{^*}\sum_{\underline{a} \bmod q} \#\mathrm{Null}_d(a_1M_1 + a_2M_2) \\ &\leq \sum_{\substack{\underline{a} \bmod q \\ (a_1,a_2,d)=1}} \#\mathrm{Null}_d(a_1M_1 + a_2M_2) \\ &= \sum_{\substack{\underline{b} \bmod d \\ (b_1,b_2,d)=1}} \#\mathrm{Null}_d(b_1M_1 + b_2M_2) \sum_{\underline{c} \bmod q/d} 1 \\ &= \left(\frac{q}{d}\right)^2 \sideset{}{^*}\sum_{\underline{b} \bmod d} \#\mathrm{Null}_d(b_1M_1 + b_2M_2) \\ &= \left(\frac{q}{d}\right)^2 S(d,d). \end{aligned} \tag{4.18}$$

For convenience, define

$$T(d) := S(d,d). \tag{4.19}$$

We firstly aim to show that $T(d)$ is multiplicative. If it is, then we will be able to consider $T(p_i)$ for some prime $p_i \mid d$, which will be far easier to work with. Let $d = d_1d_2$, such that $(d_1, d_2) = 1$. Then by (4.17), we have

$$T(d) = \sum_{\substack{\underline{b} \bmod d \\ (b_1,b_2,d)=1}} \#\mathrm{Null}_{d_1}(b_1M_1 + b_2M_2)\#\mathrm{Null}_{d_2}(b_1M_1 + b_2M_2). \tag{4.20}$$

Now, let $b_1 := u_1 + d_1v_1$, $b_2 := u_2 + d_1v_2$, and note that

$$\begin{aligned} (b_1, b_2, d) = 1 &\Leftrightarrow (b_1, b_2, d_1) = (b_1, b_2, d_2) = 1 \\ &\Leftrightarrow (u_1, u_2, d_1) = (u_1 + d_1v_1, u_2 + d_1v_2, d_2) = 1. \end{aligned}$$

Substituting this information into (4.20) gives the following:

$$\begin{aligned} T(d) &= \sideset{}{^*}\sum_{\underline{u} \bmod d_1} \#\mathrm{Null}_{d_1}(u_1M_1 + u_2M_2) \times \\ &\qquad \sum_{\substack{\underline{v} \bmod d_2 \\ (u_1+d_1v_1,\, u_2+d_1v_2,\, d_2)=1}} \#\mathrm{Null}_{d_2}([u_1 + d_1v_1]M_1 + [u_2 + d_1v_2]M_2) \\ &= \sideset{}{^*}\sum_{\underline{u} \bmod d_1} \#\mathrm{Null}_{d_1}(u_1M_1 + u_2M_2) \sum_{\substack{\underline{w} \bmod d_2 \\ (w_1,w_2,d_2)=1}} \#\mathrm{Null}_{d_2}(w_1M_1 + w_2M_2) \\ &= T(d_1)T(d_2). \end{aligned}$$

In the second step, we performed the substitution $\underline{w} = \underline{u} + d_1\underline{v}$, and then reordered the sum in terms of $\underline{w}$. We can do this because $\underline{u}$ can be treated as fixed when considering the second sum, and since $(d_1, d_2) = 1$, the map $\underline{v} \mapsto \underline{w}$ is a bijection. We have now proven that $T(d)$ is a multiplicative function. In particular, we have

$$T(d) = \prod_{i=1}^{r} T(p_i) \tag{4.21}$$

where $p_i \,|\, d$ is prime. It is therefore sufficient to consider

$$T(p) = \sideset{}{^*}\sum_{\underline{a} \bmod p} \#\{\underline{x} \bmod p : (a_1M_1 + a_2M_2)\underline{x} \equiv \underline{0} \bmod p\}, \tag{4.22}$$

where p is a prime. When $p \ll_n 1$, the right hand side is trivially $O(p^2)$. It is therefore enough to consider the case $p \gg_n 1$, where the implied constant is chosen as in the statement in Proposition 4.3. Proposition 4.3 now implies that except for $O_n(p)$ different exceptional pairs (a_1, a_2), $\mathrm{Rank}(a_1M_1 + a_2M_2) \geq n - m_p - 1$. Moreover, for the exceptional pairs we still have $\mathrm{Rank}(a_1M_1 + a_2M_2) = n - m_p - 2$. Finally, we note that if M is an integer matrix rank k over $\mathbb{F}_p$, it is easy to see that

$$\#\{\underline{x} \in \mathbb{F}_p^n : M\underline{x} = \underline{0}\} \ll p^{n-k}.$$

Applying these results to (4.22) gives us

$$T(p) \ll \sideset{}{^*}\sum_{\substack{\underline{a} \bmod p \\ \mathrm{Rank}(a_1M_1+a_2M_2)\geq n-m_p-1}} p^{m_p+1} + \sideset{}{^*}\sum_{\substack{\underline{a} \bmod p \\ \mathrm{Rank}(a_1M_1+a_2M_2)=n-m_p-2}} p^{m_p+2}$$

$$\ll p^2 \times p^{m_p+1} + p \times p^{m_p+2}$$
$$\ll p^{2+m_p+1},$$

and so

$$T(d) \ll \prod_{i=1}^{r} p_i^{2+m_{p_i}+1} = d^2 \prod_{i=1}^{r} p_i^{m_{p_i}+1}$$

by (4.21). Hence, by (4.18) - (4.19), we have

$$\begin{aligned} S(d,q) &\leq \left(\frac{q}{d}\right)^2 T(d) \\ &\ll q^2 \prod_{i=1}^{r} p_i^{m_{p_i}+1}, \end{aligned}$$

as required. □

As mentioned earlier, we aim to bound exponential sums of the form (4.2) in Chapter 8. During that process, we will need to bound the size of the set

$$N_{\underline{b},q}(M) := \{\underline{x} \in (\mathbb{Z}/q\mathbb{Z})^n \ : \ M\underline{x} \equiv \frac{q}{2}\,\underline{b} \pmod{q}\}. \tag{4.23}$$

The next lemma will help us to do this by letting us relate $N_{\underline{b},q}(M)$ to $\mathrm{Null}_q(M)$.

Lemma 4.7. *Let $q \in \mathbb{N}$ be even, $M \in M_n(\mathbb{Z}/q\mathbb{Z})$, and let $N_{\underline{b},q}(M)$ be defined as in* (4.23)*. Then for every $\underline{b} \in \{0,1\}^n$, either $N_{\underline{b},q}(M) = \emptyset$ or there exists some $\underline{y}_{\underline{b}} \in (\mathbb{Z}/q\mathbb{Z})^n$ such that*

$$N_{\underline{b},q}(M) = \underline{y}_{\underline{b}} + \mathrm{Null}_q(M).$$

Proof. If we assume that $N_{\underline{b},q}(M) \neq \emptyset$, then there must be some $\underline{y} \in N_{\underline{b},q}(M)$. By the definition of $\mathrm{Null}_q(M)$, if $\underline{y}_0 \in \mathrm{Null}_q(M)$, then $\underline{y} + \underline{y}_0 \in N_{\underline{b},q}(M)$. Hence

$$\underline{y} + \mathrm{Null}_q(M) \subset N_{\underline{b},q}(M), \tag{4.24}$$

and so $\#N_{\underline{b},q}(M) \geq \#\mathrm{Null}_q(M)$.

Likewise, we note that if $\underline{y}_1, \underline{y}_2 \in N_{\underline{b},q}(M)$, then $\underline{y}_1 - \underline{y}_2 \in \mathrm{Null}_q(M)$, and so

$\#N_{\underline{b},q}(M) \leq \#\mathrm{Null}_q(M)$. Therefore

$$\#N_{\underline{b},q}(M) = \#\mathrm{Null}_q(M) = \#(\underline{y} + \mathrm{Null}_q(M)).$$

Combining this with (4.24) gives us the result we desire. □

Chapter 5

Initial setup

In this section we will start with some initial considerations which will help us to properly set up the circle method and state our main results which will be used to prove Theorem 3.2. As stated before, the Hardy Littlewood circle method transforms the task of answering Theorem 3.2 to proving an asymptotic formula:

$$\int_0^1 \int_0^1 S(\alpha_1, \alpha_2) d\alpha_1 d\alpha_2 = C_X P^{n-6} + o(P^{n-6}). \tag{5.1}$$

Here $S(\underline{\alpha})$ is the exponential sum as defined in (3.7), and C_X denotes a product of local densities.

Remark 5.1. *In order to make some of the arguments in Chapter 8 easier to state, we will assume that* $2 \mid (\mathrm{Cont}(F), \mathrm{Cont}(G))$*, where* $\mathrm{Cont}(F)$ *is the gcd of all of its coefficients. We can assume this without loss of generality since* $F(\underline{x}) = G(\underline{x}) = 0$ *if and only if* $2F(\underline{x}) = 2G(\underline{x}) = 0$*, and so we can always opt to work with the latter forms instead if necessary.*

We will start by splitting the box $[0, 1]^2$ into a set of major arcs and minor arcs. The corresponding contribution to the integral in (5.1) over the major and minor arcs (as defined below) will give us the main contribution and the error term respectively to the asymptotic formula. We will define our major and minor arcs as follows:

For any pair (α_1, α_2), we can use a two dimensional version of Dirichlet's approximation theorem to find a simultaneous approximation $(a_1/q, a_2/q)$. In particular upon taking $Q = \left\lfloor P^{3/2} \right\rfloor$, there exists $\underline{a} = (a_1, a_2) \in \mathbb{Z}^2$ and $q \in \mathbb{N}$ s.t. $(a_1, a_2, q) = 1$, $q \leq Q$, and

$$\left|\alpha_1 - \frac{a_1}{q}\right| \leq \frac{1}{qQ^{1/2}}, \qquad \left|\alpha_2 - \frac{a_2}{q}\right| \leq \frac{1}{qQ^{1/2}}. \tag{5.2}$$

We can therefore write

$$\alpha_1 = \frac{a_1}{q} + z_1, \qquad \alpha_2 = \frac{a_2}{q} + z_2, \tag{5.3}$$

for some $|\underline{z}| := \max\{|z_1|, |z_2|\} \leq 1/qQ^{1/2}$. It is currently difficult to explain why we demand that $Q = \left\lfloor P^{3/2} \right\rfloor$ since it relates to optimising bounds that we currently do not have; we will therefore postpone discussing this until we have these bounds (see Section 12.1.1).

Now let $0 < \Delta < 1$ be some small parameter also to be chosen later, and define

$$\mathfrak{M}_{q,\underline{a}}(\Delta) := \left\{(\alpha_1, \alpha_2) \mod 1 \ : \ \left|\alpha_i - \frac{a_i}{q}\right| \leq P^{-3+\Delta}, i = 1, 2\right\}.$$

We then define the set of major arcs to be

$$\mathfrak{M} = \mathfrak{M}(\Delta) := \bigcup_{q \leq P^{\Delta}} \bigcup_{\substack{\underline{a} \bmod q \\ (\underline{a}, q) = 1}} \mathfrak{M}_{q,\underline{a}}(\Delta). \tag{5.4}$$

This union of sets is disjoint if P is sufficiently large since the individual arcs $\mathfrak{M}_{\underline{a},q}$ become too small to overlap with each other (the argument from Remark 1.2 generalises trivially to show this). Moreover, it is easy to check that $P^{-3+\Delta} < 1/qQ^{1/2}$ for any $q \leq Q$, provided that $Q < P^{3-\Delta}$. This is certainly true for our final choice $Q = P^{3/2}$ since we assumed $\Delta < 1$, and so we have that each set $\mathfrak{M}_{q,\underline{a}}$ is contained in the corresponding range from (5.2). Therefore, the major arcs give the following contribution to the integral in (5.1):

$$S_{\mathfrak{M}} := \sum_{1 \leq q \leq P^{\Delta}} \sideset{}{^*}\sum_{\underline{a} \bmod q} \int_{|\underline{z}| \leq P^{-3+\Delta}} S_{\underline{a}}(q, \underline{z}) d\underline{z}, \tag{5.5}$$

where

$$S_{\underline{a}}(q,\underline{z}) := S(\underline{a}/q+\underline{z}). \tag{5.6}$$

We then define the minor arcs to be $\mathfrak{m} = [0,1]^2\backslash\mathfrak{M}$. By construction of $\mathfrak{M}$, the individual minor arcs must therefore either have

$$P^{\Delta} < q \le Q \text{ and } |\underline{z}| < (qQ^{1/2})^{-1}, \text{ or } 1 \le q \le P^{\Delta} \text{ and } P^{-3+\Delta} < |\underline{z}| < (qQ^{1/2})^{-1}. \tag{5.7}$$

Hence, we can bound the minor arcs contribution, upon further bringing the average over $\underline{a}$ inside the integral in (5.1), by

$$S_{\mathfrak{m}} = \sum_{1\le q\le P^{\Delta}} \int_{P^{-3+\Delta}\le|\underline{z}|\le 1/qQ^{1/2}} S(q,\underline{z})d\underline{z} + \sum_{P^{\Delta}\le q\le Q} \int_{|\underline{z}|\le 1/qQ^{1/2}} S(q,\underline{z})d\underline{z}. \tag{5.8}$$

Here

$$S(q,\underline{z}) := \sideset{}{^*}\sum_{\underline{a} \bmod q} |S_{\underline{a}}(q,\underline{z})|. \tag{5.9}$$

Our techniques for dealing with the major arcs contribution are standard. Let

$$\begin{aligned} \mathfrak{S}(R) &:= \sum_{q=1}^{R} q^{-n} \sideset{}{^*}\sum_{\underline{a} \bmod q} \sum_{\underline{x} \bmod q} e_q(a_1F(\underline{x}) + a_2G(\underline{x})), \\ \mathfrak{J}(R) &:= \int_{|\underline{z}|<R} \int_{\mathbb{R}^n} \omega(\underline{x}) e(z_1F(\underline{x}) + z_2G(\underline{x}))\, d\underline{x}d\underline{z}, \end{aligned} \tag{5.10}$$

and let

$$\mathfrak{S} := \lim_{R\to\infty} \mathfrak{S}(R), \quad \mathfrak{J} = \lim_{R\to\infty} \mathfrak{J}(R), \tag{5.11}$$

denote the singular series and the corresponding singular integral, provided the limits exist. Our main major arcs estimate is the following Lemma:

Lemma 5.2. *Assume that $n - \sigma(F,G) \ge 34$, where $\sigma(F,G) := \sigma(X_{F,G})$ as defined in* (3.1)*, and assume that $\mathfrak{S}$ is absolutely convergent, satisfying*

$$\mathfrak{S}(R) = \mathfrak{S} + O_{\phi}(R^{-\phi})$$

for some $\phi > 0$. Then provided that we have $\Delta \in (0, 1/7)$,

$$S_{\mathfrak{M}} = \mathfrak{S}\mathfrak{J}P^{n-6} + O_{\phi}(P^{n-6-\delta})$$

The proof of this lemma, along with the proof of convergence of the singular series will be established in Chapter 13.

The majority of our effort will be spent in bounding the minor arcs contribution. In order to state the Proposition we aim to prove for the minor arcs, we need to further specify our choice of weight function and the point which it will centred on. Let $\underline{x}_0$ be a fixed point satisfying $|\underline{x}_0| < 1$ and

$$\text{Rank} \begin{pmatrix} \nabla F(\underline{x}_0) \\ \nabla G(\underline{x}_0) \end{pmatrix} = 2. \tag{5.12}$$

Without loss of generality, we may assume that

$$|\nabla F(\underline{x}_0) \cdot \nabla G(\underline{x}_0)| \leq C' \|\nabla F(\underline{x}_0)\|_{\mathrm{L}^2} \|\nabla G(\underline{x}_0)\|_{\mathrm{L}^2}, \tag{5.13}$$

for some $0 < C' < 1$ possibly depending on $\underline{x}_0$. We will also slightly expand our definition of the test function ω to assume it to be supported in a box $\underline{x}_0 + (-\rho, \rho)^n$, for a small parameter $\rho > 0$ to be chosen in due course. Moreover, we ask that $\omega \in \mathcal{W}_n$, where $\mathcal{W}_n$ is defined to be the set of infinitely differentiable functions $\hat{\omega} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ with compact support contained within $[-S_n, S_n]^n$ for some fixed S_n, and with the following bound to be true on its derivatives:

$$\max \left\{ \left| \frac{\partial^{j_1+\cdots+j_n}}{\partial x_1^{j_1} \cdots \partial x_n^{j_n}} \hat{\omega}(\underline{x}) \right| \mid \underline{x} \in \mathbb{R}, j_1 + \cdots + j_n = j \right\} \ll_{j,n} 1 \tag{5.14}$$

for every $j \geq 0$. A satisfactory bound for the minor arcs will be produced by the following proposition, which we aim to prove:

Proposition 5.3. *Let F, G be a system of two cubic forms satisfying $n \geq 39$. Let $m_\infty(F, G) = -1$, and let $\omega \in \mathcal{W}_n$, where $\underline{x}_0$ satisfies* (5.13)*. Then there exists some $\delta = \delta(\Delta) > 0$ and some $\rho_0 > 0$, such that for any $0 < \Delta < 1/7$ and for any $0 < \rho < \rho_0$, we have*

$$S_{\mathfrak{m}} = O_{n,\rho,\Delta,||F||,||G||}(P^{n-6-\delta}).$$

Here, $m_\infty(F, G)$ is as defined in (3.3)*.*

A major part of the rest of this work will be dedicated to proving Proposition

5.3, which will ultimately be achieved in Section 12. Before we move on, it will be desirable to obtain a consequence of our choice of ω and $\underline{x}_0$, akin to the conditions [21, (2.15)-(2.16)]. This will be our aim in lemma 5.4 below, which will be useful in setting up a two dimensional van der Corput differencing argument in Section 7 and in particular, in the proof of Lemma 7.3. We choose the vectors $\underline{e}_1'$ and $\underline{e}_2'$ to be a basis for the span of the two dimensional vector space $\{\nabla F(\underline{x}_0), \nabla G(\underline{x}_0)\}$, chosen in the following way:

$$\underline{e}_1' := \frac{\nabla F(\underline{x}_0)}{||\nabla F(\underline{x}_0)||}, \quad \underline{e}_2' := \frac{\nabla G(\underline{x}_0) - \gamma \underline{e}_1'}{\gamma_1}, \tag{5.15}$$

where $\gamma = \nabla G(\underline{x}_0) \cdot \underline{e}_1'$, and $\gamma_1 = \|\nabla G(\underline{x}_0) - \gamma \underline{e}_1'\|$ is a non-zero constant by (5.13). $\underline{e}_2'$ is chosen so that the Gram-Schmidt procedure works.

Lemma 5.4. *Let F and G be cubic forms and ω be a compactly supported function supported in $\underline{x}_0 + (-\rho, \rho)^n$ satisfying* (5.14)*, where $\underline{x}_0$ satisfies* (5.13)*. Then there exist constants $M_1, M_2 > 0$ such that*

$$\min_{\underline{x} \in \mathrm{Supp}(P\omega)} |\nabla F(\underline{x}) \cdot \underline{e}_1'| \geq M_1 P^2, \quad \min_{\underline{x} \in \mathrm{Supp}(P\omega)} |\nabla G(\underline{x}) \cdot \underline{e}_2'| \geq M_1 P^2, \tag{5.16}$$

$$\max_{\underline{x} \in \mathrm{Supp}(P\omega)} \{|\nabla F(\underline{x}) \cdot \underline{e}_2'|\} \leq \rho M_2 P^2, \quad \max_{\underline{x} \in \mathrm{Supp}(P\omega)} \{|\nabla G(\underline{x}) \cdot \underline{e}_1'|\} \leq M_2 P^2. \tag{5.17}$$

Furthermore, there exists some $0 < \rho_0 \leq 1$ such that if $\rho \leq \rho_0$, then M_1 and M_2 depend only on F, G, and our choice of $\underline{x}_0$ (in particular M_1 and M_2 do not depend on ρ).

Proof. A key in the proof here will be the following bound, which is an easy consequence of the Mean Value Theorem: Given any $\underline{x} \in \mathrm{Supp}(P\omega)$, we have

$$\|\nabla F(\underline{x}) - \nabla F(P\underline{x}_0)\| \ll_{||F||} \rho P^2 \text{ and } \|\nabla G(\underline{x}) - \nabla G(P\underline{x}_0)\| \ll_{||G||} \rho P^2. \tag{5.18}$$

Let us first prove that the conditions for $\nabla F(\underline{x})$ in (5.16) - (5.17) are met. The key here are the conditions (5.12) and (5.13). Clearly, using (5.18) we have

$$\nabla F(\underline{x}) \cdot \underline{e}_1' = (\nabla F(\underline{x}) - \nabla F(P\underline{x}_0)) \cdot \underline{e}_1' + \nabla F(P\underline{x}_0) \cdot \underline{e}_1'$$

$$
\begin{aligned}
&= (\nabla F(\underline{x}) - \nabla F(P\underline{x}_0)) \cdot \underline{e}_1' + P^2 \nabla F(\underline{x}_0) \cdot \nabla F(\underline{x}_0)/\|\nabla F(\underline{x}_0)\| \\
&= (\nabla F(\underline{x}) - \nabla F(P\underline{x}_0)) \cdot \underline{e}_1' + P^2 \|\nabla F(\underline{x}_0)\| \\
&\geq (1 - O(\rho))P^2 \|\nabla F(\underline{x}_0)\| \\
&\geq M_{F,1} P^2
\end{aligned}
$$

for some $M_{F,1} > 0$ which is independent of ρ, provided that ρ is chosen to be small enough. Similarly, we may also assure that

$$
|\nabla G(\underline{x}) \cdot \nabla G(\underline{x}_0)| \geq (1 - O(\rho))P^2 \|\nabla G(\underline{x}_0)\|^2. \tag{5.19}
$$

In both of these equations, the implied constants only depend on $\|F\|, ||G||$ and n. This will be a feature of all implied constants appearing in this proof. On the other hand, since $\nabla F(\underline{x}_0) = \|\nabla F(\underline{x}_0)\| \, \underline{e}_1'$ is orthogonal to $\underline{e}_2'$, we have

$$
|\nabla F(\underline{x}) \cdot \underline{e}_2'| = |(\nabla F(\underline{x}) - P^2 \nabla F(\underline{x}_0)) \cdot \underline{e}_2'| \leq \|(\nabla F(\underline{x}) - \nabla F(P\underline{x}_0))\| \ll_{\|F\|} \rho P^2 \tag{5.20}
$$

by (5.18). In other words, there is some $M_{F,2} > 0$ independent of ρ such that

$$
|\nabla F(\underline{x}) \cdot \underline{e}_2'| \leq M_{F,2} \, \rho P^2.
$$

To deal with the inequalities concerning G, we use (5.13), which hands us a constant $0 < C' < 1$ satisfying

$$
\begin{aligned}
\gamma \|\nabla F(\underline{x}_0)\| = |\nabla F(\underline{x}_0) \cdot \nabla G(\underline{x}_0)| &\leq C' \|\nabla F(\underline{x}_0)\|_{L_2} \|\nabla G(\underline{x}_0)\|_{L_2} \\
&\leq C' \|\nabla F(\underline{x}_0)\| \|\nabla G(\underline{x}_0)\|.
\end{aligned} \tag{5.21}
$$

Therefore, for any $\underline{x} \in \mathrm{Supp}(P\omega)$, by (5.18) and (5.21), we have that

$$
\begin{aligned}
|\nabla F(\underline{x}_0) \cdot \nabla G(\underline{x})| &\leq |\nabla F(\underline{x}_0) \cdot \nabla G(P\underline{x}_0)| + \\
&\qquad |\nabla F(\underline{x}_0) \cdot (\nabla G(\underline{x}) - \nabla G(P\underline{x}_0))| \\
&\leq C' P^2 \|\nabla G(\underline{x}_0)\| \|\nabla F(\underline{x}_0)\| + O_{\|G\|}(\rho) P^2 \|\nabla F(\underline{x}_0)\|
\end{aligned}
$$

Hence (since $\|\nabla G(\underline{x}_0)\| > 0$ is a constant), provided that the support ρ is sufficiently

small, we may choose some $0 < C'' < 1$ independent of ρ such that

$$|\nabla F(\underline{x}_0) \cdot \nabla G(\underline{x})| \leq C'' P^2 \|\nabla F(\underline{x}_0)\| \|\nabla G(\underline{x}_0)\|. \tag{5.22}$$

Thus, for any $\underline{x} \in \mathrm{Supp}(P\omega)$,

$$\begin{aligned} |\nabla G(\underline{x}) \cdot (\nabla G(\underline{x}_0) - \gamma \underline{e}_1')| &= |\nabla G(\underline{x}) \cdot \nabla G(\underline{x}_0) - \\ &\qquad \gamma \|\nabla F(\underline{x}_0)\|^{-1} \nabla G(\underline{x}) \cdot \nabla F(\underline{x}_0)| \\ &\geq (1 - O(\rho) - C'C'') P^2 \|\nabla G(\underline{x}_0)\|^2, \end{aligned}$$

where we have used (5.21) to bound γ by $C' \|\nabla G(\underline{x}_0)\|$, as well as (5.22) and (5.19). Hence provided that the support ρ is chosen to be sufficiently small, there is some $M_{G,1} > 0$ such that

$$|\nabla G(\underline{x}) \cdot \underline{e}_2'| = \gamma_1^{-1} |\nabla G(\underline{x}) \cdot (\nabla G(\underline{x}_0) - \gamma \underline{e}_1')| \geq M_{G,1} P^2.$$

Hence, upon taking

$$M_1 := \min\{M_{F,1}, M_{G,1}\},$$

we conclude that (5.16) is true. Finally, (5.22) also hands us:

$$|\nabla G(\underline{x}) \cdot \underline{e}_1'| = \|\nabla F(\underline{x}_0)\|^{-1} |\nabla F(\underline{x}_0) \cdot \nabla G(\underline{x})| \leq C'' P^2 \|G(\underline{x}_0)\|, \tag{5.23}$$

for any $\underline{x} \in \mathrm{Supp}(P\omega)$. Therefore, upon setting $M_{2,G} := C'' \|G(\underline{x}_0)\|$, and taking

$$M_2 := \max\{M_{F,2}, M_{G,2}\},$$

we are now able to verify (5.17). Furthermore, there is some $\rho_0 > 1$, such that M_1 and M_2 are independent of ρ provided that $\rho \leq \rho_0$. This concludes the proof of the lemma. □

Chapter 6

Weyl Differencing

In order to make this thesis more accessible to those less familiar with the circle method, we will begin by bounding our exponential sums using Weyl differencing. We would normally cover van der Corput differencing first, but the arguments from Weyl differencing are simpler than those coming from van der Corput. Throughout this chapter, we will work in the more general setting of polynomials (as opposed to forms).

Weyl differencing in the context of the circle method has been studied extensively by many people, perhaps most notably by Birch in his landmark paper in 1961 [1], and almost all results which improve upon Birch's theorem method use Weyl differencing in some form or another. It is therefore unsurprising that we will need several bounds which use Weyl differencing, but unlike in [1], they will serve as complimentary bounds to the more powerful ones coming from van der Corput differencing and Poisson summation.

The main idea that goes into any differencing method is quite simple: We aim to use it to bound $S_{\underline{a}}(q, \underline{z})$ (see (5.6)) from above by considering the square of its absolute value, which will ultimately lead to us bounding $S_{\underline{a}}(q, \underline{z})$ by an exponential sum with polynomials of degree $d - 1$. In the case of Weyl differencing, we then repeat this process until we end up with linear polynomials. From here, there are

standard methods to bound exponential sums with linear polynomials from above non-trivially, but we will come to that later.

In order to prove Proposition 5.3, we will need a bound which uses Weyl differencing twice (to go from cubics to linear polynomials), as well as two bounds which come from applying variations of van der Corput differencing once, followed by a single application of Weyl differencing on the resulting quadratic exponential sum. The derivation of these bounds are very similar to each other; we will therefore only detail the process of using Weyl differencing to bound the quadratic exponential sums coming from applying van der Corput to a cubic (see (7.13)).

In the case of the former: The topic of performing Weyl differencing repeatedly on a system of forms has already been covered extensively by Lee in the context of function fields [20]. The Weyl differencing arguments that are used in his paper do not rely on being in a function fields setting, and so we may freely invoke the results in [20, Section 3]. In particular, upon setting $d = 3$ and $R = 2$, an application of [20, Lemma 3.7] gives us

$$|S(\underline{a}/q + \underline{z})| \ll P^{n+\varepsilon}\Big(P^{-4} + q^2|\underline{z}|^2 + q^2P^{-6} + q^{-1}\min\{1, \frac{1}{|\underline{z}|P^3}\}\Big)^{(n-\sigma'-1)/16},$$

where

$$\sigma' = \sigma'(F^{(0)}, G^{(0)}) := \dim\{\underline{x} \in \mathbb{P}^{n-1}_{\mathbb{C}} \ : \ \mathrm{Rank}\begin{pmatrix}\nabla F^{(0)}(\underline{x})\\ \nabla G^{(0)}(\underline{x})\end{pmatrix} < 2\}, \tag{6.1}$$

and $F^{(0)}$, $G^{(0)}$ are defined to be the cubic components of F and G respectively. However, we may use Proposition 4.4 to conclude that $\sigma' \leq \sigma(F^{(0)}, G^{(0)}) + 1$. Hence, by applying two Weyl differencing steps to $|S(\underline{a}/q + \underline{z})|$, we arrive at the following:

Proposition 6.1 (Weyl/Weyl)**.** *Let F, G be cubic polynomials such that*

$$\|F^{(0)}\|, \|G^{(0)}\| \asymp 1,$$

and $\sigma(F^{(0)}, G^{(0)}) = \sigma$. *Then:*

$$|S(\underline{a}/q + \underline{z})| \ll P^{n+\varepsilon}\Big(P^{-4} + q^2|\underline{z}|^2 + q^2P^{-6} + q^{-1}\min\{1, \frac{1}{|\underline{z}|P^3}\}\Big)^{(n-\sigma-2)/16}.$$

We now aim to bound the exponential sum,

$$T(q, \underline{z}) := \sum_{\underline{a}}^{*} \sum_{\underline{x}\in\mathbb{Z}^n} \omega(\underline{x}/P)e([a_1/q + z_1]F(\underline{x}) + [a_2/q + z_2]G(\underline{x}))$$

that we get after performing van der Corput differencing once. In this case, F and G are quadratic polynomials such that $\|F^{(0)}\|, \|G^{(0)}\| \ll H$, for some $1 \leq H \leq P$. For the remainder of the chapter, we will work through a less general version of the argument used by Lee in [20, Section 3].

We start by considering the exponential sum

$$T(\underline{a}, q, \underline{z}) := \sum_{\underline{x}\in\mathbb{Z}^n} \omega(\underline{x}/p)e([a_1/q + z_1]F(\underline{x}) + [a_2/q + z_2]G(\underline{x})) \tag{6.2}$$

where $\omega \in \mathcal{W}_n \cup \{\chi\}$ where χ is the characteristic function on $(0, 1]^n$, and F, G are quadratic polynomials. In this case we cannot directly use [20, Lemma 3.7] to bound $T(\underline{a}, q, \underline{z})$ because we do not necessarily have $\|F^{(0)}\|, \|G^{(0)}\| \asymp 1$. Fortunately, the majority of the arguments used in [20, Section 3] do not rely on this assumption, enabling us to follow the same procedure up to a few minor adjustments.

To begin, let

$$F^{(0)}(\underline{x}) = \sum_{j_1,j_2=1}^{n} b_{j_1,j_2}x_{j_1}x_{j_2}, \quad G^{(0)}(\underline{x}) = \sum_{j_1,j_2=1}^{n} c_{j_1,j_2}x_{j_1}x_{j_2} \tag{6.3}$$

where $b_{i,j} = b_{j,i}$ and $c_{i,j} = c_{j,i}$, and let

$$|F^{(0)}| := \sum_{j_1,j_2} |b_{j_1,j_2}|, \quad |G^{(0)}| := \sum_{j_1,j_2} |c_{j_1,j_2}|, \quad \lambda = \lambda_{F,G} := 2\max\{|F^{(0)}|, |G^{(0)}|\}. \tag{6.4}$$

Then, since we need $\underline{x} \in P\mathrm{Supp}(\omega)$ in order for $\omega(\underline{x}/P) \neq 0$:

$$\begin{aligned}|T(\underline{a}, q, \underline{z})|^2 &= \Big|\sum_{|\underline{x}|\ll P} \omega(\underline{x}/P)e([a_1/q + z_1]F(\underline{x}) + [a_2/q + z_2]G(\underline{x}))\Big|^2 \\ &\leq \sum_{|\underline{y}|\ll P} \Big|\sum_{|\underline{x}|\in\mathbb{Z}^n} \omega(\underline{x})\overline{\omega(\underline{y}/P)}e\Big([a_1/q + z_1](F(\underline{x}) - F(\underline{y})) +\end{aligned}$$

$$[a_2/q + z_2](G(\underline{x}) - G(\underline{y}))\Big)\Big|.$$

We have chosen ω so that $\mathrm{Supp}(\omega) \subset \underline{x}_0 + (-\rho, \rho)^n$, where $0 < \rho < 1$ and $|\underline{x}_0| < 1$ are fixed, and so we can replace $\sum_{\underline{y} \ll P}$ with $\sum_{\underline{y} < P}$, provided that we choose ρ to be sufficiently small. Since we are summing over the entire integer lattice, we may replace $\underline{x}$ with $\underline{x} + \underline{y}$. We also note that ω is a real valued function, and so upon setting

$$F(\underline{x}, \underline{y}) := F(\underline{x} + \underline{y}) - F(\underline{x}), \quad G(\underline{x}, \underline{y}) := G(\underline{x} + \underline{y}) - G(\underline{x}),$$
$$\omega_{\underline{y},P}(\underline{x}) := \omega((\underline{x} + \underline{y})/P)\omega(\underline{y}/P),$$

we have

$$|T(\underline{a}, q, \underline{z})|^2 \leq \sum_{|\underline{y}|<P} \Big| \sum_{\underline{x} \in \mathbb{Z}^n} \omega_{\underline{y},P}(\underline{x}) e([a_1/q + z_1]F(\underline{x}, \underline{y})) + [a_2/q + z_2]G(\underline{x}, \underline{y}))\Big|. \quad (6.5)$$

The upshot of all of this is that $F(\underline{x}, \underline{y}), G(\underline{x}, \underline{y})$ are bilinear, and so if we set $\alpha_j = a_j/q + z_j$ for $j \in \{1, 2\}$, then there must be some linear $L_{\underline{\alpha},j}(\underline{y}), \Phi_{\underline{\alpha}}(\underline{y})$ such that

$$\alpha_1 F(\underline{x}, \underline{y}) + \alpha_2 G(\underline{x}, \underline{y}) = \sum_{i=1}^{n} x_i L_{\underline{\alpha},j}(\underline{y}) + \Phi_{\underline{\alpha}}(\underline{y}).$$

It is also easy to check that

$$L_{\underline{\alpha},j}(\underline{y}) = \alpha_1 B_{1,j}(\underline{y}) + \alpha_2 B_{2,j}(\underline{y}) \quad (6.6)$$

where

$$B_{1,j}(\underline{y}) := 2\sum_{i=1}^{n} b_{i,j} y_i, \quad B_{2,j}(\underline{y}) := 2\sum_{i=1}^{n} c_{i,j} y_i. \quad (6.7)$$

Hence, by (6.5), we have

$$|T(\underline{a}, q, \underline{z})|^2 \leq \sum_{|\underline{y}|<P} \Big| \sum_{\underline{x} \in \mathbb{Z}^n} \omega_{\underline{y},P}(\underline{x}) e\Big(\sum_{j=1}^{n} x_j L_{\underline{\alpha},j}(\underline{y})\Big)\Big|.$$

If $\omega = \chi$, then we immediately have

$$|T(\underline{a}, q, \underline{z})|^2 \leq \sum_{|\underline{y}|<P} \prod_{j=1}^{n} \Big| \sum_{|x_i| \ll P} e(x_j L_{\underline{\alpha},j}(\underline{y}))\Big| \quad (6.8)$$

$$\ll \sum_{|\underline{y}|<P} \prod_{j=1}^{n} \min\{P, \langle L_{\underline{\alpha},j}(\underline{y})\rangle^{-1}\}, \tag{6.9}$$

where $\langle x\rangle$ is the distance of x to the nearest integer. This is due to the sums in (6.8) being geometric series. Likewise, when $\omega_{\underline{y},P} \in \mathcal{W}_n$, we can reach the same expression using partial summation. (6.9) is clearly at its worst when $\langle L_{\underline{\alpha},j}(\underline{y})\rangle$ is relatively close to zero, so we expect that any further analysis will take this into account. In the next Lemma, we see that this intuition is correct.

Lemma 6.2. $|T(\underline{a}, q, \underline{z})|^2 \ll P^n (\log P)^n \# N(\underline{\alpha}, P)$ *where*

$$N(\underline{\alpha}, P) := \{|\underline{y}| < P \ : \ \langle L_{\underline{\alpha},j}(\underline{y})\rangle < P^{-1} \ \forall j \leq n\}.$$

Proof. We will adapt an argument of Davenport [9, Lemma 13.2] to see this. We claim that for any integers $r_1, \cdots, r_n$ s.t. $0 \leq r_j < P$, there can be at most $\# N(\underline{\alpha}, P)$ values of $\underline{y} \ll P$ for which the system of inequalities

$$\frac{r_j}{P} \leq \{L_{\underline{\alpha},j}(\underline{y})\} < \frac{r_j + 1}{P} \tag{6.10}$$

is true, where $\{x\}$ is the fractional part of x. Indeed if we let $\underline{y}_1, \underline{y}_2 \ll P$ (not necessarily distinct) be such that (6.10) is true, then by linearity of the $L_{\underline{\alpha},j}$'s in $\underline{y}$, $\underline{y}_1 - \underline{y}_2 \in N(\underline{\alpha}, P)$. Hence we cannot have more than $\# N(\underline{\alpha}, P)$ distinct $\underline{y}_i$'s satisfying (6.10) as each pair $(\underline{y}_1, \underline{y}_i)$ corresponds to an element of $N(\underline{\alpha}, P)$.

Hence, if we let

$$N(\underline{\alpha}, P, \underline{r}) := \left\{|\underline{y}| < P \ : \ \frac{r_j}{P} \leq \{L_{\underline{\alpha},j}(\underline{y})\} < \frac{r_j + 1}{P} \ \forall j \leq n\right\}$$

then

$$\# N(\underline{\alpha}, P, \underline{r}) \leq N(\underline{\alpha}, P)$$

for every $\underline{r} \in \mathbb{Z}^n_{\geq 0}$, $|\underline{r}| < P$. Now, if (6.10) is true, then

$$\frac{r_j}{P} \leq \langle L_{\underline{\alpha},j}(\underline{y})\rangle < \frac{r_j + 1}{P}$$

if $r_j < P/2$ and

$$\frac{P - r_j - 1}{P} \leq \langle L_{\underline{\alpha},j}(\underline{y})\rangle < \frac{P - r_j}{P}$$

otherwise. Therefore, we have

$$\begin{aligned}\sum_{|\underline{y}|<P}\prod_{j=1}^{n}\min\{P, \langle L_{\underline{\alpha},j}(\underline{y})\rangle^{-1}\} &\leq \sum_{r_1=0}^{P-1}\cdots\sum_{r_n=0}^{P-1}\#N(\underline{\alpha}, P, \underline{r})\times \\ &\qquad \prod_{j=1}^{n}\min\left\{P, \max\left\{\frac{P}{r_j}, \frac{P}{P - r_j - 1}\right\}\right\} \\ &\leq P^n\#N(\underline{\alpha}, P)\times \\ &\qquad \sum_{r_1=0}^{P-1}\cdots\sum_{r_n=0}^{P-1}\prod_{j=1}^{n}\min\left\{1, \max\left\{\frac{1}{r_j}, \frac{1}{P - r_j - 1}\right\}\right\} \\ &\ll P^n(\log P)^n\#N(\underline{\alpha}, P),\end{aligned}$$

and so by (6.9), we arrive at $|T(\underline{a}, q, \underline{z})|^2 \ll P^n(\log P)^n\#N(\underline{\alpha}, P)$ as claimed. □

Our next task is to find a good bound on the size of $\#N(\underline{\alpha}, P)$. It would be nice if we could replace $\langle L_{\underline{\alpha},j}(\underline{y})\rangle < P^{-1}$ in the definition of $N(\underline{\alpha}, P)$ by $\langle L_{\underline{\alpha},j}(\underline{y})\rangle < K$, for some small K, because this would likely force some non-trivial condition on the $B_{i,j}(\underline{y})$'s. For example, we might hope for something like $a_1B_{1,j}(\underline{y}) + a_2B_{2,j}(\underline{y}) = 0$ to be necessary if K was chosen to be sufficiently small (recall $\alpha = \underline{a}/q + \underline{z}$). This is ultimately not the condition we need, but it illustrates the point that if we could replace P^{-1} by something smaller in $\#N(\underline{\alpha}, P)$, then it may lead to some non-trivial restriction on $\underline{y}$. This would in turn lead us to a non-trivial bound on the size of the corresponding set (where $\langle L_{\underline{\alpha},j}(\underline{y})\rangle < P^{-1}$ is replaced with $\langle L_{\underline{\alpha},j}(\underline{y})\rangle < K$).

To this end, we will state an important auxiliary lemma of Davenport, and then use it to bound $\#N(\underline{\alpha}, P)$ by such a set.

Lemma 6.3. *Let L be a real symmetric $n \times n$ matrix, and let*

$$N(Z) := \{\underline{u} \in \mathbb{Z}^n \ : \ |\underline{u}| \leq PZ, \ \langle (L\underline{u})_j\rangle < P^{-1}Z \ \forall j \leq n\}.$$

Then, if $0 < Z_1 \leq Z_2 \leq 1$*, we have*

$$\#N(Z_2) \ll_c \left(\frac{Z_2}{Z_1}\right)^n \#N(Z_1).$$

This lemma is a specific case ($a = P$) of Lemma 12.6 from Davenport's book: *Analytic Methods for Diophantine Equations and Inequalities*, and the proof of it can naturally be found there (see [9][Lemma 12.6]). We have slightly rephrased the lemma to better fit in with our notation, but it is easy to check that this is indeed equivalent to Davenport's lemma.

We will now use Lemma 6.3 to replace $N(\underline{\alpha}, P)$ in Lemma 6.2 by a set whose size we can bound more easily. We start by noting that if we set $(L\underline{u})j = L_{\underline{\alpha},j}(\underline{u})$, then

$$N(\underline{\alpha}, P) = \{\underline{u} \in \mathbb{Z}^n \,:\, |\underline{u}| \leq P,\ \langle L_{\underline{\alpha},j}(\underline{u})\rangle < P^{-1}\ \forall j \leq n\} = N(1).$$

We also have that the matrix constructed out of the coefficients of the $L_{\underline{\alpha},j}$'s (see (6.6)) is symmetric, since $b_{i,j} = b_{j,i}$, $c_{i,j} = c_{j,i}$ in (6.7) (this is due to how we defined the $b_{i,j}$'s and $c_{i,j}$'s in (6.3)).

Therefore, Lemma 6.3 gives us

$$\#N(\underline{\alpha}, P) \ll_d Z^{-n} \#N(Z), \tag{6.11}$$

for any $Z < 1$, where

$$N(Z) = N_Z(\underline{\alpha}, P) := \{\underline{y} \in \mathbb{Z}^n \,:\, |\underline{y}| < ZP,\ \langle L_{\underline{\alpha},j}(\underline{y})\rangle < ZP^{-1}\ \forall\, j \leq n\}. \tag{6.12}$$

We have now bounded $\#N(\underline{\alpha}, P)$ in terms of a set whose $\underline{y}$'s run over a smaller box of whatever size is convenient for us. This is a crucial result for us, so we will state this as a lemma.

Lemma 6.4. *For any* $0 < Z \leq 1$*, we have*

$$|T(\underline{a}, q, \underline{z})|^2 \ll P^{n+\varepsilon} Z^{-n} \#N_Z(\underline{\alpha}, P), \tag{6.13}$$

We now aim to choose Z optimally to minimise the right-hand side of the bound in

Lemma 6.4. To this end, we will introduce the following specific case of [20, Lemma 3.6]:

Lemma 6.5. *Let $N > 0$, $\underline{\alpha} = \underline{a}/q + \underline{z}$, $|\underline{z}| < (4qN)^{-1}$, $(a_1, a_2, q) = 1$. Let*

$$M := \begin{pmatrix} m_{1,1} & \cdots & m_{1,n} \\ m_{2,1} & \cdots & m_{2,n} \end{pmatrix}$$

be a $2 \times n$ matrix such that $m_{i,j} \in \mathbb{Z}^2$, $|m_{i,j}| < N$, for every i, j, and assume that $\langle \alpha_1 m_{1,j} + \alpha_2 m_{2,j} \rangle < \tilde{Q}^{-1}$ for some $\tilde{Q} \geq 4q$. Then

$$a_1 m_{1,j} + a_2 m_{2,j} \equiv 0 \mod q$$

for every $j \in \{1, \cdots, n\}$, and

$$q \mid \det M_{i,j} := \det \begin{pmatrix} m_{1,i} & m_{1,j} \\ m_{2,i} & m_{2,j} \end{pmatrix}$$

for every $(i, j) \in \{1, \cdots, n\}^2$, $i \neq j$. Furthermore, if in addition either $N^2 < q/2$ or $|\underline{z}| > 2N(q\tilde{Q})^{-1}$, then $\mathrm{Rank}_{\mathbb{Q}}(M) < 2$.

Proof. The argument for the first assertion is very simple and works in a similar way to [13, Lemma 2.3]. Using the triangle inequality repeatedly:

$$\begin{aligned} \left\langle \frac{a_1 m_{1,j} + a_2 m_{2,j}}{q} \right\rangle &\leq \langle \alpha_1 m_{1,j} + \alpha_2 m_{2,j} \rangle + \langle z_1 m_{2,j} \rangle + \langle z_2 m_{2,j} \rangle \\ &< \tilde{Q}^{-1} + 2N|z_i| \\ &\leq (4q)^{-1} + 2N(4qN)^{-1} = \frac{3}{4} q^{-1} < q^{-1}. \end{aligned}$$

This implies that

$$a_1 m_{1,j} + a_2 m_{2,j} \equiv 0 \mod q \text{ for every } j \in \{1, \cdots, n\}, \tag{6.14}$$

and since $(a_1, a_2, q) = 1$, we must have that

$$\mathrm{Rank}(M \bmod q) < 2.$$

In particular, if we define $M_{i,j}$ to be the minor of M constructed out of the i-th and j-th columns, then this implies that $q|\det(M_{i,j})$ for every $(i,j) \in \{1,\cdots,n\}^2$, $i \neq j$. We now aim to verify the second half of the lemma.

Assume for a contradiction that $\text{Rank}_{\mathbb{Q}}(M) = 2$. Then there must be a 2×2 minor of M, $M_{i,j}$, such that $\det(M_{i,j}) \neq 0$. For the sake of notational simplicity, we will assume that this is the leading minor of M, $M_{1,2}$. In particular, this implies that

$$m_{1,1}m_{2,2} - m_{1,2}m_{2,1} \neq 0.$$

Since $q \mid \det(M_{1,2})$, we must have that $q \mid m_{1,1}m_{2,2} - m_{1,2}m_{2,1}$. Now, if $N^2 < q/2$, then

$$|m_{1,1}m_{2,2} - m_{1,2}m_{2,1}| \leq |m_{1,1}| \cdot |m_{2,2}| - |m_{1,2}| \cdot |m_{2,1}| < 2N^2 < q.$$

Therefore we must have that $m_{1,1}m_{2,2} - m_{1,2}m_{2,1} = 0$ since $m_{i,j} \in \mathbb{Z}$ for every i,j, contradicting our assumption that $\det(M_{1,2}) \neq 0$. If instead $|\underline{z}| \geq 2N(q\tilde{Q})^{-1}$, then we use a slightly more complex argument:

Firstly we note that

$$|\underline{z}| \cdot |m_{1,1}m_{2,2} - m_{1,2}m_{2,1}| \geq 4N(q\tilde{Q})^{-1}q = 2N\tilde{Q}^{-1}, \tag{6.15}$$

since we are assuming that $\det(M_{1,2}) \neq 0$. We also have that $|m_{i,j}z_i| = \langle m_{i,j}z_i \rangle$ for every i,j since

$$|m_{i,j}z_i| \leq (4qN)^{-1}|m_{i,j}| < (4q)^{-1} \leq \frac{1}{4}.$$

In particular, this implies that

$$|m_{1,j_1}z_1| + |m_{2,j_2}z_2| < \frac{1}{2} \tag{6.16}$$

Hence, for any $j_1, j_2 \in \{1,\cdots,n\}$

$$\begin{aligned} m_{1,j_1}m_{2,j_2}|\underline{z}| &\leq m_{2,j_2}|m_{1,j_1}z_1| + m_{1,j_1}|m_{2,j_2}z_2| \\ &\leq N(|m_{1,j_1}z_1| + |m_{2,j_2}z_2|) \\ &= N(\langle m_{1,j_1}z_1 \rangle + \langle m_{2,j_2}z_2 \rangle) \end{aligned}$$

$$= N\langle m_{1,j_1}z_1 + m_{2,j_2}z_2\rangle \tag{6.17}$$

Since (6.16) implies that

$$|m_{1,j_1}z_1| + |m_{2,j_2}z_2| = \langle m_{1,j_1}z_1\rangle + \langle m_{2,j_2}z_2\rangle = \langle m_{1,j_1}z_1 + m_{2,j_2}z_2\rangle.$$

Hence, by (6.17) and the triangle inequality

$$\begin{aligned} m_{1,j_1}m_{2,j_2}|\underline{z}| &\leq N\Big(\langle m_{1,j_1}\alpha_1 + m_{2,j_2}\alpha_2\rangle + \Big\langle\frac{m_{1,j_1}a_1 + m_{2,j_2}a_2}{q}\Big\rangle\Big) \\ &< N\tilde{Q}^{-1}. \end{aligned}$$

The final inequality is due to the assumption in the lemma that $\langle\alpha_1 m_{1,j_1} + \alpha_1 m_{1,j_1}\rangle < \tilde{Q}^{-1}$ and (6.14). However, this implies that

$$\begin{aligned} |\underline{z}| \cdot |m_{1,1}m_{2,2} - m_{1,2}m_{2,1}| &\leq |\underline{z}| \cdot |m_{1,1}m_{2,2}| + |\underline{z}| \cdot |m_{1,2}m_{2,1}| \\ &< N\tilde{Q}^{-1} + N\tilde{Q}^{-1} = 2N\tilde{Q}^{-1}, \end{aligned}$$

contradicting (6.15). Hence we may also conclude that $\mathrm{Rank}_{\mathbb{Q}}(M) < 2$ in the case that $|\underline{z}| \geq 2N(q\tilde{Q})^{-1}$. □

We now aim to choose Z in a way which will enable us to use Lemma 6.5 to bound $\#N_Z(\underline{\alpha}, P)$. In preparation for this, we will define

$$N := \lambda PZ, \quad \tilde{Q} := Z^{-1}P, \tag{6.18}$$

and let

$$M(\underline{y}) = M := \begin{pmatrix} B_{1,1}(\underline{y}) & \cdots B_{1,n}(\underline{y}) \\ B_{2,1}(\underline{y}) & \cdots B_{2,n}(\underline{y}) \end{pmatrix} \tag{6.19}$$

We defined $N, \tilde{Q}, M$ in this way as a set-up for showing that the first part of Lemma 6.5 is true for any $\underline{y} \in N_Z(\underline{\alpha}, P)$. Indeed, upon recalling (6.4), we see that for any $\underline{y} \in N_Z(\underline{\alpha}, P)$,

$$|m_{i,j}| = |B_{i,j}(\underline{y})| \leq \max\{|F^{(0)}|, |G^{(0)}|\} \times \max_{|\underline{y}| \in N_Z(\underline{\alpha}, P)} |\underline{y}| \leq \lambda \times PZ = N,$$

since the $B_{i,j}$'s are linear in $\underline{y}$. We also automatically have that

$$\langle \alpha_1 m_{1,j} + \alpha_2 m_{2,j} \rangle = \langle \alpha_1 B_{1,j}(\underline{y}) + \alpha_2 B_{2,j}(\underline{y}) \rangle < \tilde{Q}^{-1}$$

for every j, by the definitions of $N_Z(\underline{\alpha}, P)$ and $\tilde{Q}$ (see (6.12), (6.18)). However, in order for $|\underline{z}| < (4qN)^{-1}$ and $\tilde{Q} \geq 4q$, we need to choose Z appropriately. In particular, we will choose

$$Z \leq (4q|\underline{z}|\lambda P)^{-1}, \quad Z \leq P/4q, \quad 0 < Z \leq 1, \tag{6.20}$$

where λ is as in (6.4). If we choose Z in this way, then by (6.18)

$$N \leq \lambda P(4q|\underline{z}|\lambda P)^{-1} = (4q|\underline{z}|)^{-1}. \quad \Leftrightarrow \quad |\underline{z}| \leq (4qN)^{-1},$$

and

$$\tilde{Q} = Z^{-1}P \geq (P/4q)^{-1}P = 4q.$$

The final condition of (6.20) ensures that Lemma 6.4 is not violated. Therefore, provided that we define $N, \tilde{Q}, M$ as in (6.18) - (6.19), and provided that (6.20) is true, then every $\underline{y} \in N_Z(\underline{\alpha}, P)$ then $q \mid \det M_{i,j}$ for every $(i,j) \in \{1, \cdots, n\}$, $i \neq j$.

We now need to choose Z so that either $N^2 < q/2$ or $|\underline{z}| > 2N(q\tilde{Q})^{-1}$, so that we can guarantee that $\mathrm{Rank}_{\mathbb{Q}}(M) < 2$ (as this condition will ultimately lead us to a non-trivial bound for $N_Z(\underline{\alpha}, P)$). To this end, we will choose

$$Z^2 \leq \max\left\{ \frac{q}{2\lambda^2 P^2}, \frac{q|\underline{z}|}{2\lambda} \right\}. \tag{6.21}$$

If $Z^2 \leq q/(2\lambda^2 P^2)$, then

$$N^2 = \lambda^2 P^2 Z^2 \leq \lambda^2 P^2 \times q(2\lambda^2 P^2)^{-1} = q/2.$$

Alternatively, if $Z^2 \leq q|\underline{z}|/(2\lambda)$, then

$$2N(q\tilde{Q})^{-1} = 2(\lambda P Z) \times q^{-1}(ZP^{-1}) = 2\lambda Z^2 q^{-1} \leq 2\lambda q^{-1} \times q|\underline{z}|(2\lambda)^{-1} = |\underline{z}|.$$

Hence, as long as we define $N, \tilde{Q}, M$ as in (6.18) - (6.19), and choose Z so that (6.20)-(6.21) are true, then we have that $\mathrm{Rank}_{\mathbb{Q}}(M(\underline{y})) < 2$, for every $\underline{y} \in N_Z(\underline{\alpha}, P)$.

In particular, this implies that

$$N_Z(\underline{\alpha}, P) \subset \{\underline{y} \in (\mathbb{Z} \cap [-ZP, ZP])^n \ : \ \mathrm{Rank}_{\mathbb{Q}}(M) < 2\}.$$

From here, it is easy to check that

$$\partial_j F^{(0)}(\underline{y}) = B_{1,j}(\underline{y}), \quad \partial_j G^{(0)}(\underline{y}) = B_{2,j}(\underline{y})$$

and so we may replace $\mathrm{Rank}_{\mathbb{Z}}(M) < 2$ with the following:

$$N_Z(\underline{\alpha}, P) \subset T(ZP) := \{\underline{y} \in (\mathbb{Z} \cap [-ZP, ZP])^n \ : \ \mathrm{Rank}_{\mathbb{Q}} \begin{pmatrix} \nabla F^{(0)}(\underline{y}) \\ \nabla G^{(0)}(\underline{y}) \end{pmatrix} < 2\}. \quad (6.22)$$

But, by the definition of $\sigma'(F, G)$ (see (6.1)) and Proposition 4.4, we have that

$$\#T(R) \ll R^{\sigma'(F^{(0)}, G^{(0)})+1} \ll R^{\sigma(F^{(0)}, G^{(0)})+2}.$$

Hence upon letting $\sigma(F^{(0)}, G^{(0)}) := \sigma$, (6.13) and (6.22) give us

$$|T(\underline{a}, q, \underline{z})|^2 \ll Z^{-n+\sigma+2} P^{n+\varepsilon+\sigma+2} \quad (6.23)$$

provided that $Z \geq P^{-1}$. The bound is trivially true for $Z < P^{-1}$. In order to minimise our bound for $|T(\underline{a}, q, \underline{z})|$, we should choose Z to be as large as possible, whilst respecting the constraints on it that enabled us to reach (6.23) (see (6.20)-(6.21)). In particular, will choose

$$Z \asymp \min\left\{1, (\lambda^2 q^2 |\underline{z}|^2 P^2)^{-1}, \frac{P^2}{q^2}, \max\left\{\frac{q}{\lambda^2 P^2}, \frac{q|\underline{z}|}{\lambda}\right\}\right\}^{1/2}.$$

Substituting this into (6.23) gives us

$$\begin{aligned} |T(\underline{a}, q, \underline{z})|^2 &\ll P^{n+\sigma+1+\varepsilon}\Big(1 + \lambda^2 q^2 P^2 |\underline{z}|^2 + q^2 P^{-2} + q^{-1}\lambda^2 \min\{P^2, \frac{1}{\lambda|\underline{z}|}\}\Big)^{(n-\sigma-2)/2} \\ &= P^{2n+\varepsilon}\Big(P^{-2} + \lambda^2 q^2 |\underline{z}|^2 + q^2 P^{-4} + q^{-1}\lambda^2 \min\{1, \frac{1}{\lambda|\underline{z}|P^2}\}\Big)^{(n-\sigma-2)/2} \end{aligned}$$

Therefore, we have the following:

Proposition 6.6 (van der Corput/Weyl)**.** *Let F, G be quadratic polynomials such*

that

$$\|F^{(0)}\|, \|G^{(0)}\| \leq H,$$

and let $\sigma := \sigma(F^{(0)}, G^{(0)})$. *Then:*

$$|T(\underline{a}, q, \underline{z})| \ll P^{n+\varepsilon}\Big(P^{-2} + q^2H^2|\underline{z}|^2 + q^2P^{-4} + q^{-1}H^2 \min\{1, \frac{1}{H|\underline{z}|P^2}\}\Big)^{(n-\sigma-2)/4}.$$

Chapter 7

Van der Corput differencing

In this Chapter, we will use the more powerful van der Corput differencing to bound $S_{\underline{a}}(q, \underline{z})$ from above by a quadratic exponential sum. We will introduce the topic by beginning with the simpler *pointwise* van der Corput differencing used in [3] before attempting to generalise the differencing arguments used in [28] to attain a bound which also takes advantage of averaging over the both $\underline{z}$ integrals. In both cases, we will innovate on the standard differencing approach in order to introduce a path to attaining Kloosterman refinement.

7.1 Pointwise van der Corput

In order to understand how Weyl and the standard van der Corput differencing differ from each other, we will firstly recall (6.5): After performing Weyl differencing once, we ended up with the bound

$$|T(\underline{a}, q, \underline{z})|^2 \leq \sum_{|\underline{y}|<P} \Big| \sum_{\underline{x}\in\mathbb{Z}^n} \omega_{\underline{y},P}(\underline{x}) e([a_1/q + z_1]F(\underline{x}, \underline{y})) + [a_2/q + z_2]G(\underline{x}, \underline{y}))\Big|.$$

In particular, we note that our $\underline{y}$ sum goes up to P. Our aim with van der Corput differencing is to let the $\underline{y}$ sum go up to some H, where $1 \leq H \ll P$ can be chosen freely. In particular, as long as we can find a way to choose this H optimally, we

should be able to attain a better bound than the one we found via Weyl differencing unless the optimal value for H happens to be P.

For convenience, we will set

$$\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}) := (a_1/q + z_1)F(\underline{x}) + (a_2/q + z_2)G(\underline{x}), \tag{7.1}$$

where F and G are cubic forms. Since $\underline{x}$ is summed over all of $\mathbb{Z}^n$, we can replace $\underline{x}$ with $\underline{x} + \underline{h}$, for any $\underline{h} \in \mathbb{Z}^n$, giving

$$S(q,\underline{z}) = \sum_{\underline{a}}{}^{*}\Big|\sum_{\underline{x}\in\mathbb{Z}^n}\omega((\underline{x}+\underline{h})/P)e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}+\underline{h}))\Big|, \tag{7.2}$$

where $S(q,\underline{z})$ is as defined in (5.9). Let $\mathcal{H} \subset \mathbb{Z}^n$ be a set of lattice points (which we may choose freely). In the case of pointwise van der Corput differencing, we can just take $\mathcal{H}$ to be the set of lattice points $\underline{h}$ such that $|\underline{h}| < H$, but we will not specify this in the arguments that follow since we will need a different choice of $\mathcal{H}$ when we come to averaged van der Corput differencing later. Then, (7.2) and the Cauchy-Schwarz inequality gives the following

$$\begin{aligned}
\#\mathcal{H}S(q,\underline{z}) &= \sum_{\underline{a}}{}^{*}\Big|\sum_{\underline{h}\in\mathcal{H}}\sum_{\underline{x}\in\mathbb{Z}^n}\omega((\underline{x}+\underline{h})/P)e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}+\underline{h}))\Big| \\
&\le \sum_{\underline{a}}{}^{*}\sum_{\underline{x}\in\mathbb{Z}^n}\Big|\sum_{\underline{h}\in\mathcal{H}}\omega((\underline{x}+\underline{h})/P)e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}+\underline{h}))\Big| \\
&\le \Big(\sum_{\underline{a}}{}^{*}\sum_{|\underline{x}|<2P}1\Big)^{1/2}\Big(\sum_{\underline{a}}{}^{*}\sum_{\underline{x}\in\mathbb{Z}^n}\Big|\sum_{\underline{h}\in\mathcal{H}}\omega((\underline{x}+\underline{h})/P)e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}+\underline{h}))\Big|^2\Big)^{1/2} \\
&\ll qP^{n/2}\Big(\sum_{\underline{a}}{}^{*}\sum_{\underline{x}\in\mathbb{Z}^n}\sum_{\underline{h}_1,\underline{h}_2\in\mathcal{H}}\omega((\underline{x}+\underline{h}_1)/P)\overline{\omega((\underline{x}+\underline{h}_2)/P)} \\
&\qquad\qquad e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}+\underline{h}_1))\overline{e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}+\underline{h}_2))}\Big)^{1/2}.
\end{aligned}$$

The key difference between this and the standard van der Corput differencing process is the introduction of the $\underline{a}$ sum in the Cauch-Schwarz step. In particular, this enables us to bring the $\underline{a}$ sum inside of the bracket in the final step which in turn gives us a path to Kloosterman refinement. We still need to write $S(q,\underline{z})$ in terms of a quadratic exponential sum however, so we will come back to Kloosterman refinement later.

Set $\underline{y} := \underline{x} + \underline{h}_2$, $\underline{h} = \underline{h}_1 - \underline{h}_2$ and recall that we defined ω to be a real weight function. Therefore, after setting

$$N(\underline{h}) := \#\{\underline{h}_2 - \underline{h}_1 = \underline{h} : \underline{h}_1, \underline{h}_2 \in \mathcal{H}\}, \text{ and } \omega_{\underline{h}}(\underline{x}) := \omega(\underline{x} + P^{-1}\underline{h})\omega(\underline{x}), \tag{7.3}$$

we get

$$|S(q,\underline{z})|^2 \ll \#\mathcal{H}^{-2}q^2P^n \sideset{}{^*}\sum_{\underline{a}} \sum_{\underline{y}\in\mathbb{Z}^n}\sum_{\underline{h}\in\mathcal{H}} N(\underline{h})\omega_{\underline{h}}(\underline{y}/P)e(\hat{F}_{\underline{a},q,\underline{z}}(\underline{y}+\underline{h}) - \hat{F}_{\underline{a},q,\underline{z}}(\underline{y})).$$

Recall that $\hat{F}_{\underline{a},q,\underline{z}}(\underline{x}) = (a_1/q + z_1)F(\underline{x}) + (a_2/q + z_2)G(\underline{x})$. Therefore if we set $F_{\underline{h}}$ and $G_{\underline{h}}$ be the differenced polynomials

$$F_{\underline{h}}(\underline{y}) := F(\underline{y}+\underline{h}) - F(\underline{y}), \quad G_{\underline{h}}(\underline{y}) := G(\underline{y}+\underline{h}) - G(\underline{y}),$$

we have

$$\hat{F}_{\underline{a},q,\underline{z}}(\underline{y}+\underline{h}) - \hat{F}_{\underline{a},q,\underline{z}}(\underline{y}) = (a_1/q + z_1)F_{\underline{h}}(\underline{y}) + (a_2/q + z_2)G_{\underline{h}}(\underline{y}).$$

Hence

$$|S(q,\underline{z})|^2 \ll \#\mathcal{H}^{-2}P^nq^2 \sum_{\underline{h}\in\mathcal{H}} N(\underline{h})T_{\underline{h}}(q,\underline{z}), \tag{7.4}$$

where

$$T_{\underline{h}}(q,\underline{z}) := \sideset{}{^*}\sum_{\underline{a} \bmod q} \sum_{\underline{y}\in\mathbb{Z}^n} \omega_{\underline{h}}(\underline{y}/P)e((a_1/q + z_1)F_{\underline{h}}(\underline{y}) + (a_2/q + z_2)G_{\underline{h}}(\underline{y})) \tag{7.5}$$

denote the corresponding exponential sum for the system of quadratic polynomials $F_{\underline{h}}$ and $G_{\underline{h}}$. Note that the top form of $F_{\underline{h}}$, $F_{\underline{h}}^{(0)}$, is precisely (3.10). Finally, by noting that $N(\underline{h}) \le \#\mathcal{H} = H^n$, we arrive at the following:

Lemma 7.1. *For any $1 \le H \ll P$, for any fixed choice of $\underline{z} \in [0,1]^2$, we have*

$$|S(q,\underline{z})| \ll H^{-n/2}P^{n/2}q\Big(\sum_{\underline{h}\ll H} |T_{\underline{h}}(q,\underline{z})|\Big)^{1/2}.$$

This bound will be useful to us when $t := |\underline{z}|$ is small, say of size $P^{-3-\Delta}$, since it is wasteful to use *averaged* van der Corput differencing in this case. We will now set up averaged van der Corput differencing, which will be a key in proving Proposition

5.3.

7.2 Averaged van der Corput

Throughout this section, we will work on generalising the differencing method used in [10] and [21] to work in the context of two forms. $\underline{x}_0$ will denote a fixed point satisfying $|\underline{x}_0| < 1$ in $\underline{x}_0 \in \mathrm{Supp}(\omega)$, where $\mathrm{Supp}(\omega)$ is contained in the set $\underline{x}_0 + (-\rho, \rho)^n$. Likewise, F and G will be cubic polynomials whose leading forms satisfy (5.16) and (5.17) for a fixed orthonomal set of vectors $\underline{e}'_1, \underline{e}'_2$ (see (5.15)). Let

$$\{\underline{e}'_1, ..., \underline{e}'_n\}, \tag{7.6}$$

denote an extended orthonormal basis of $\mathbb{R}^n$. We will begin our effort to bound the sum

$$\sum_{P^{\Delta} \leq q \leq Q} \int_{P^{-3-\Delta} \leq |\underline{z}| \leq 1/qQ^{1/2}} S(q, \underline{z})\, d\underline{z}, \tag{7.7}$$

where $S(q, \underline{z}) = \sum^{*}_{\underline{a} \bmod q} |S_{\underline{a}}(q, \underline{z})|$ is as defined in (5.9). As in the previous section, let $1 \leq H \ll P$ be a parameter to be chosen later. Typically, H will be chosen as a small power of P, so it is safe to further assume $H \log P \ll P$. Also, let $\varepsilon > 0$ be an arbitrarily small absolute constant to be chosen at the end. Note that the implied constants will be allowed to depend on the choice of ε after it is introduced into our bounds. As is standard ([28] for example), we start by splitting the integral over $\underline{z}$ above as a sum over $O(P^{\varepsilon})$ *dyadic intervals* of the form $[t, 2t]$ where $P^{-3+\Delta} \leq t \leq 1/(qQ^{1/2})$. For convenience, given $t \in \mathbb{R}^2_{>0}$, we will set

$$I(q, t) := \int_{t \leq |\underline{z}| \leq 2t} S(q, \underline{z})\, d\underline{z}.$$

Analogous to [10] and [21, Section 3], for a fixed value of $P^{-3-\Delta} < t < 1/qQ^{1/2}$ we choose two sets T_1, T_2 of cardinality $O(1 + tHP^2)$ such that

$$\begin{aligned}\{\underline{z} : t \leq |\underline{z}| \leq 2t\} \subseteq \bigcup_{\underline{\tau} \in T_1 \times T_2} \Big[\tau_1 - (HP^2)^{-1}, \tau_1 + (HP^2)^{-1}\Big] \times \\ \Big[\tau_2 - (HP^2)^{-1}, \tau_2 + (HP^2)^{-1}\Big] \\ \subseteq \{\underline{z} : t - (HP^2)^{-1} \leq |\underline{z}| \leq 2t + (HP^2)^{-1}\}.\end{aligned} \tag{7.8}$$

Thus, an application of Cauchy-Schwarz further gives

$$I(q,t) \ll ((HP^2)^{-1} + t) \sum_{\underline{\tau} \in \underline{T}} \mathcal{M}_q(\underline{\tau}, H)^{1/2}, \tag{7.9}$$

where

$$\begin{aligned}\mathcal{M}_q(\underline{\tau}, H) :&= \int_{\underline{\tau} - (HP^2)^{-1}}^{\underline{\tau} + (HP^2)^{-1}} |S(q, \underline{z})|^2 \, d\underline{z} \\ &\ll \int_{\mathbb{R}^2} \exp(-H^2P^4[(\tau_1 - z_1)^2 + (\tau_2 - z_2)^2]) |S(q, \underline{z})|^2 \, d\underline{z}.\end{aligned} \tag{7.10}$$

Here we have used $\underline{T} := T_1 \times T_2$ and $\int_{\underline{\tau} - (HP^2)^{-1}}^{\underline{\tau} + (HP^2)^{-1}}$ to denote the integral

$$\int_{(\tau_1 - (HP^2)^{-1}, \tau_1 + (HP^2)^{-1}) \times (\tau_2 - (HP^2)^{-1}, \tau_2 + (HP^2)^{-1})}$$

in order to simplify the notation. After an inspection of the right hand side of (7.8), it is easy to see that

$$\int_{P^{-3-\Delta} \leq |\underline{z}| \leq 1/qQ^{1/2}} S(q, \underline{z}) \, d\underline{z} \ll \sum_t ((HP^2)^{-1} + t) \sum_{\underline{\tau} \in \underline{T}} \mathcal{M}_q(\underline{\tau}, H)^{1/2},$$

where the sum over t runs over $O_\varepsilon(P^\varepsilon)$ choices satisfying

$$P^{-3-\Delta} \leq t \leq 1/(qQ). \tag{7.11}$$

Note that the choice of the parameter H will ultimately depend on t. For now, we will assume t to be fixed.

We are therefore first led to find a bound for $|S(q, \underline{z})|^2$ using van der Corput differencing. We may now use the same arguments as those from Section 7.1 to arrive at

the following:

$$|S(q,\underline{z})|^2 \ll \#\mathcal{H}^{-2}P^n q^2 \sum_{\underline{h}\in\mathcal{H}} N(\underline{h})T_{\underline{h}}(q,\underline{z}), \tag{7.12}$$

where $\mathcal{H} \subset \mathbb{Z}^n$ is a set of lattice points to be chosen later, and

$$T_{\underline{h}}(q,\underline{z}) := \sum_{\underline{a} \bmod q}^{*} \sum_{\underline{y}\in\mathbb{Z}^n} \omega_{\underline{h}}(\underline{y}/P)e((a_1/q+z_1)F_{\underline{h}}(\underline{y}) + (a_2/q+z_2)G_{\underline{h}}(\underline{y})) \tag{7.13}$$

denotes the corresponding exponential sum for the system of quadratic polynomials $F_{\underline{h}}$ and $G_{\underline{h}}$ (this is a restating of (7.4) and (7.5)).

Therefore by (7.9), (7.10), and (7.12), we have shown the following:

Lemma 7.2. *For any $1 \le H \le P$, $\mathcal{H} \subset \mathbb{Z}^n$, and t satisfying (7.11) we have*

$$I(q,t) \ll (HP^2)^{-1}\#\mathcal{H}^{-1}P^{n/2}q \times \sum_{\underline{\tau}\in\underline{T}} \Big(\sum_{\underline{h}\in\mathcal{H}} N(\underline{h}) \int_{\mathbb{R}^2} \exp(-H^2P^4[(\tau_1-z_1)^2+(\tau_2-z_2)^2])T_{\underline{h}}(q,\underline{z})\, d\underline{z}\Big)^{1/2}. \tag{7.14}$$

Since we intend to develop a two dimensional version of averaged van der Corput differencing, we intend to choose $\mathcal{H}$ to be a set of size $O(P^2H^{n-2})$ and then use averaging over z_1 and z_2 to show that for all but $O((H\log(P))^n)$ of $\underline{h} \in \mathcal{H}$, the value of the averaged integral $\mathcal{M}_q(\underline{\tau}, H)$ defined in (7.10) is negligible. This will enable us to 'win' an extra factor of P/H in our final estimate for (7.7) when compared to pointwise van der Corput differencing.

Our choice of $\mathcal{H}$ will be informed by the following lemma:

Lemma 7.3. *For any $\underline{h} \in \mathbb{R}^n$, any $1 \le H \le P$, any fixed $\underline{\tau}$ and any $N > 0$,*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-H^2P^4[(\tau_1-z_1)^2+(\tau_2-z_2)^2])T_{\underline{h}}(q,\underline{z})\, d\underline{z} \ll_N P^{-N},$$

provided that $\underline{h} = \sum_{i=1}^n h_i'\underline{e}_i'$ satisfies the following condition:

$$H\mathcal{L} \ll |h_1'| \ll P \quad \text{or} \quad H\mathcal{L} \ll |h_2'| \ll P, \quad |h_i'| < H \text{ for } i \in \{3,\cdots,n\}, \tag{7.15}$$

where $\mathcal{L} = \log(P)$, $\{\underline{e}_1', ..., \underline{e}_n'\}$ denote the basis chosen in (7.6) and the implied constants only depend on $n, \|F\|$ and $\|G\|$.

Proof. We start by rewriting the expression in the lemma and then integrating:

$$\int_{\mathbb{R}^2} \exp(-H^2P^4[(\tau_1 - z_1)^2 + (\tau_2 - z_2)^2])T_{\underline{h}}(q, \underline{z})\, d\underline{z}$$
$$= \sum_{\underline{y}\in\mathbb{Z}^n} \sum_{\underline{a}}{}^{*} \omega_{\underline{h}}(\underline{y}/P) e_q(a_1 F_{\underline{h}}(\underline{y}) + a_2 G_{\underline{h}}(\underline{y})) J(\underline{h}, \underline{y})$$

where

$$J(\underline{h}, \underline{y}) = \int_{\mathbb{R}^2} \exp(-H^2P^4[(\tau_1 - z_1)^2 + (\tau_2 - z_2)^2]) e(z_1 F_{\underline{h}}(\underline{y}) + z_2 G_{\underline{h}}(\underline{y})) d\underline{z} \quad (7.16)$$

and $e_q(x) := e^{2\pi i x/q}$. We may separate the two integrals over $\underline{z}$ and integrate them to get

$$J(\underline{h}, \underline{y}) = \frac{\pi}{H^2P^4} \exp\Big(- \frac{\pi^2}{H^2P^4}\Big(|F_{\underline{h}}(\underline{y})|^2 + |G_{\underline{h}}(\underline{y})|^2\Big)\Big) e(-\tau_1 F_{\underline{h}}(\underline{y}) - \tau_2 G_{\underline{h}}(\underline{y})).$$

We note that if either $|F_{\underline{h}}(\underline{y})|$ or $|G_{\underline{h}}(\underline{y})|$ are $\gg HP^2\mathcal{L}$, then trivially bounding everything in J from above gives:

$$\sum_{\underline{y}\in\mathbb{Z}^n} \sum_{\underline{a} \bmod q}^{*} \omega_{\underline{h}}(\underline{y}/P) e_q(a_1 F_{\underline{h}}(\underline{y}) + a_2 G_{\underline{h}}(\underline{y})) J(\underline{h}, \underline{y}) \ll P^n q^2 \frac{1}{H^2P^4} \exp(-m\mathcal{L}^2)$$
$$\ll_N P^{-N}$$

for some constant $m > 0$. Therefore it is sufficient to show that there exist constants $0 < c_1, c_2 < 1$ such that for every $\underline{h} \in \mathbb{R}^n$ with

$$H\mathcal{L} \ll |h_1'| < c_1 P \quad \text{or} \quad H\mathcal{L} \ll |h_2'| < c_2 P, \quad |h_i'| < H \text{ for } i \in \{3, \cdots, n\}, \quad (7.17)$$

we have

$$|F_{\underline{h}}(\underline{y})| \gg HP^2\mathcal{L} \quad \text{or} \quad |G_{\underline{h}}(\underline{y})| \gg HP^2\mathcal{L}, \quad (7.18)$$

where $\underline{h}' = (h_1', \cdots, h_n')$ is defined by

$$\underline{h} = \sum_{i=1}^{n} h_i \underline{e}_i = \sum_{i=1}^{n} h_i' \underline{e}_i'. \quad (7.19)$$

We will rewrite $F_{\underline{h}}$ as follows:

$$F_{\underline{h}}(\underline{y}) = \nabla F(\underline{y}) \cdot \underline{h} + \underline{h}^t \mathcal{H}_F(\underline{y}) \underline{h} + F_{\underline{h}}^{(2)}$$

where $F_{\underline{h}}^{(2)}$ is the constant part of $F_{\underline{h}}$ and $\mathcal{H}_F(\underline{y})$ is the Hessian of F evaluated at $\underline{y}$. Now for $\underline{h}$ satisfying (7.17), we have

$$\begin{aligned} F_{\underline{h}}(\underline{y}) &= \nabla F(\underline{y}) \cdot \underline{h} + \Big(\sum h_i' \underline{e}_i' \Big)^t \mathcal{H}_F(\underline{y}) \Big(\sum h_i' \underline{e}_i' \Big) + F_{\underline{h}}^{(2)} \\ &= \nabla F(\underline{y}) \cdot \underline{h} + F_{\underline{h}}^{(2)} + O(|h_1'|^2 P) + O(|h_2'|^2 P) + O(HP^2), \end{aligned} \tag{7.20}$$

where $F_{\underline{h}}^{(2)}$ is a cubic polynomial in $\underline{h}$, and the implied constants depend only on $\|F\|$, $\|G\|$ and n. Note that

$$F_{\underline{h}}^{(2)} = O(|h_1'|^3) + O(|h_2'|^3) + O(H^3),$$

and so we may simplify (7.20) to

$$F_{\underline{h}}(\underline{y}) = \nabla F(\underline{y}) \cdot \underline{h} + O(|h_1'|^2 P) + O(|h_2'|^2 P) + O(HP^2), \tag{7.21}$$

since $H, |h_1'|, |h_2'| < P$. We also write $\underline{h} = h_1' \underline{e}_1' + ... + h_n' \underline{e}_n'$ and invoke (5.16) and (5.17) to further get that for all $\underline{y} \in \mathrm{Supp}(P\omega)$ we have

$$|\nabla F(\underline{y}) \cdot \underline{h}| \geq |h_1'| M_1 P^2 + O(\rho |h_2'| P^2) + O(HP^2),$$

and so we get

$$|F_{\underline{h}}(\underline{y})| \geq M_1 |h_1'| P^2 + O(\rho |h_2'| P^2) + O(|h_1'|^2 P) + O(|h_2'|^2 P) + O(HP^2), \tag{7.22}$$

by (7.21). For now, let us focus on the case $|h_2'| \ll \rho^{-1/2} |h_1'|$. In this case, we must have that h_1' satisfies (7.17). Furthermore, upon choosing $c_1 \leq \rho^2$ and by (7.17), we have

$$\begin{aligned} \rho |h_2'| P^2 &\ll \rho^{1/2} |h_1'| P^2 \\ |h_1'|^2 P &\leq c_1 |h_1'| P^2 \leq \rho^2 |h_1'| P^2 \\ |h_2'|^2 P &\ll \rho^{-1} |h_1'|^2 P \leq \rho^{-1} c_1 |h_1'| P^2 \leq \rho |h_1'| P^2 \\ HP^2 &\ll |h_1'| P^2 \mathcal{L}^{-1} \ll \rho |h_1'| P^2. \end{aligned}$$

Hence, we may simplify (7.22) to

$$|F_{\underline{h}}(\underline{y})| \geq M_1|h_1'|P^2 + O(\rho^{1/2}|h_1'|P^2) \gg |h_1'|P^2 \gg HP^2\mathcal{L},$$

provided that ρ is chosen to be sufficiently small with respect to M_1.

It now remains to study the case $|h_1'| \ll \rho^{1/2}|h_2'|$. In this case, we instead have that h_2' must satisfy the bound in (7.17). We now apply the same process used to obtain (7.21) to $G_{\underline{h}}(\underline{y})$ to obtain

$$G_{\underline{h}}(\underline{y}) = \nabla G(\underline{y}) \cdot \underline{h} + O(|h_1'|^2P) + O(|h_2'|^2P) + O(HP^2) \tag{7.23}$$

where the implied constants again depend only on n, $\|F\|$ and $\|G\|$. Note again that

$$\nabla G(\underline{y}) \cdot \underline{h} = h_1'\nabla G(\underline{y}) \cdot \underline{e}_1' + h_2'\nabla G(\underline{y}) \cdot \underline{e}_2' + O(HP^2).$$

Combining this with (7.23), and applying (5.16) - (5.17) gives

$$|G_{\underline{h}}(\underline{y})| \geq M_1|h_2'|P^2 + O(|h_1'|P^2) + O(|h_1'|^2P) + O(|h_2'|^2P) + O(HP^2). \tag{7.24}$$

We now aim to simplify (7.24). Using the assumption that $|h_1'| \ll \rho^{1/2}|h_2'|$, the fact that $|h_2'|$ must obey (7.17) in this case, and setting $c_2 \leq \rho$ we have

$$\begin{aligned}
|h_1'|P^2 &\ll \rho^{1/2}|h_2'|P^2 \\
|h_1'|^2P &\ll \rho|h_2'|^2P \leq \rho c_2|h_2'|P^2 \leq \rho^2|h_2'|P^2 \\
|h_2'|^2P &\leq c_2|h_2'|P^2 \leq \rho|h_1'|P^2 \\
HP^2 &\ll |h_1'|P^2\mathcal{L}^{-1} \ll \rho|h_1'|P^2.
\end{aligned}$$

Hence

$$|G_{\underline{h}}(\underline{y})| \geq M_1|h_2'|P^2 + O(\rho^{1/2}|h_2'|P^2) \gg |h_2'|P^2 \gg HP^2\mathcal{L},$$

as long as ρ is chosen small enough.

□

The lemma above leads to the following natural choice for $\mathcal{H}$:

$$\mathcal{H} := \{\underline{h} \in \mathbb{Z}^n : 0 \leq h_1' < c_1 P,\ 0 \leq h_2' < c_2 P,\ 0 \leq h_i' < H \text{ for } i \in \{3, \cdots, n\}\}, \tag{7.25}$$

where c_1 and c_2 are the implied constants arising in (7.15). Essentially, $\mathcal{H}$ is chosen to be the collection of lattice points inside of a fixed n dimensional cuboid, B_P, centred at the origin, with volume $\mathrm{Vol}(B_P) = c_1 c_2 P^2 H^{n-2}$. The sides of the cuboid are in the direction of the basis vectors $\{\underline{e}_1', \cdots, \underline{e}_n'\}$. We now claim that

$$P^2 H^{n-2} \ll \#\mathcal{H} \ll P^2 H^{n-2}. \tag{7.26}$$

This follows very easily from the following asymptotic formula for a general cuboid B with side lengths $l_1, \cdots, l_n$. It is easy to see that

$$\#\{\mathbb{Z}^n \cap B\} = \mathrm{Vol}(B) + \sum_{i=1}^{n} O(\prod_{j \neq i} l_j).$$

The error comes from estimating the $n-1$ dimensional boundary of B. In our case $l_1 = c_1 P$,$l_2 = c_2 P$, $l_i = H$ for $i \geq 3$, which leads to (7.26). Now, the reason why we picked $\mathcal{H}$ as in (7.25) is so that we can use the bound that we found in Lemma (7.3). In particular, we can now show the following:

Lemma 7.4. *Let $1 \leq H \leq P$ and let*

$$\tilde{\mathcal{H}} := \{\underline{h} \in \mathbb{Z}^n : |\underline{h}| \ll H\mathcal{L}\}.$$

Then for any $1 \leq H \leq P$, any $1 \leq N$, and any $t > 0$ such that (7.11) holds, we have

$$I(q,t) \ll H^{-n/2+1} (\log P)^{1/2} P^{n/2-1} q((HP^2)^{-1} + t)^2 \Big(\sum_{\underline{h} \in \tilde{\mathcal{H}}} \max_{\underline{z}} |T_{\underline{h}}(q, \underline{z})| \Big)^{1/2} + O_N(P^{-N}),$$

where the maximum over $\underline{z}$ is taken over the set

$$t - (HP^2)^{-1} \leq |\underline{z}| \leq 2t + (HP^2)^{-1}. \tag{7.27}$$

Proof. Let $\mathcal{H}$ be as in (7.7). Then we use the decomposition $\mathcal{H} = \tilde{\mathcal{H}} \cup \mathcal{H} \backslash \tilde{\mathcal{H}}$. By

construction,

$$\mathcal{H}\backslash\tilde{\mathcal{H}} = \{\underline{h} \in \mathbb{Z}^n : H\mathcal{L} \ll |h_1'| < c_1 P \text{ or } \ H\mathcal{L} \ll |h_2'| < c_2 P, \\ |h_i'| < H, \text{ for } i \in \{3, \cdots, n\}\}.$$

Furthermore, note that for any fixed $\underline{h}$, $N(\underline{h})$ as defined in (7.3) satisfies the bound

$$N(\underline{h}) \ll \#\mathcal{H} \ll P^2 H^{n-2}. \tag{7.28}$$

Therefore by Lemma 7.3,

$$\#\mathcal{H}^{-1}\left(\sum_{\underline{h}\in\mathcal{H}\backslash\tilde{\mathcal{H}}} N(\underline{h})\int_{\mathbb{R}^2}\exp(-H^2P^4[(\tau_1-z_1)^2+(\tau_2-z_2)^2])T_{\underline{h}}(q,\underline{z})d\underline{z}\right)^{1/2} \ll P^{-N}$$

Further combining with the bounds $q \le Q \le P^{3/2}$ and $\#\underline{T} \ll (1+tHP^2)^2 \ll P^6$, which arises from using crude bounds $t \le 1$ and $1 \le H \le P$, we may bound the contribution from the sum over $\underline{h} \in \mathcal{H} \setminus \tilde{\mathcal{H}}$ in (7.14) as follows:

$$((HP^2)^{-1}+t)P^{n/2}q\#\mathcal{H}^{-1} \times \\ \sum_{\underline{\tau}\in\underline{T}}\Big(\sum_{\underline{h}\in\mathcal{H}\backslash\tilde{\mathcal{H}}} N(\underline{h})\int_{\mathbb{R}^2}\exp(-H^2P^4[(\tau_1-z_1)^2+(\tau_2-z_2)^2])T_{\underline{h}}(q,\underline{z})\,d\underline{z}\Big)^{1/2} \\ \ll_N P^{-2+n/2+3/2-N} \ll_N P^{(n-1)/2-N} \ll_{n,N} P^{-N},$$

as N is allowed to be arbitrarily large. Therefore, combining this with Lemma 7.2, we get

$$I(q,\underline{t}) \ll ((HP^2)^{-1}+t)\#\mathcal{H}^{-1/2}P^{n/2}q \\ \times\sum_{\tau\in T}\Big(\sum_{\underline{h}\in\tilde{\mathcal{H}}}\int_{\mathbb{R}^2}\exp(-H^2P^4[(\tau_1-z_1)^2+(\tau_2-z_2)^2])T_{\underline{h}}(q,\underline{z})\,d\underline{z}\Big)^{1/2} \\ + O_{n,N}(P^{-N}). \tag{7.29}$$

Further note that for a fixed τ and for any z satisfying $|z-\tau| \ge HP^2\mathcal{L}$ we have the

following decay of the function in the integrand:

$$\exp(-H^2P^4(\tau-z)^2) \ll \frac{\exp(-\mathcal{L}^2/2)}{|z-\tau|^2+1} \ll_N \frac{P^{-N}}{|z-\tau|^2+1}. \tag{7.30}$$

Thus, in the same vein as before, using bound (7.30) in (7.29) we may obtain

$$I(q,\underline{t}) \ll ((HP^2)^{-1}+t)\#\mathcal{H}^{-1/2}P^{n/2}q\sum_{\underline{\tau}\in\underline{T}}\Big(\sum_{\underline{h}\in\tilde{\mathcal{H}}}\int_{\underline{\tau}-(HP^2)^{-1}\mathcal{L}}^{\underline{\tau}+(HP^2)^{-1}\mathcal{L}}|T_{\underline{h}}(q,\underline{z})|\,d\underline{z}\Big)^{1/2} + O_{n,N}(P^{-N}).$$

The lemma now follows after using (7.26) to estimate $\#\mathcal{H}$, using the estimate $\#\underline{T} = O((1+tHP^2)^2)$, and (7.8) which allows us to take the maximum over all possible $\underline{z}$ appearing in the expression. □

Since H is arbitrary, we may re-label $H\mathcal{L}$ as H at the expense of a factor of size at most $O_\varepsilon(P^\varepsilon)$ we can now conclude the following

Lemma 7.5. *For any $1 \le H \ll P$, any $0<\varepsilon<1$, any $\underline{t}$ satisfying* (7.11) *and any $N \ge 1$ we have*

$$I(q,t) \ll_{\varepsilon,n,N} H^{-n/2+1}P^{n/2-1+\varepsilon}q((HP^2)^{-1}+t)^2\Big(\max_{|\underline{z}|}\sum_{|\underline{h}|\ll H}|T_{\underline{h}}(q,\underline{z})|\Big)^{1/2} + P^{-N},$$

where the maximum over $\underline{z}$ is taken over the set

$$t - P^\varepsilon(HP^2)^{-1} \le |\underline{z}| \le 2t + P^\varepsilon(HP^2)^{-1}. \tag{7.31}$$

Chapter 8

Quadratic Exponential Sums: Initial Consideration

The van der Corput technique used in Section 7 leads us to consider quadratic exponential sums $T_{\underline{h}}(q, \underline{z})$ (see (7.13)) for a family of differenced quadratic forms $F_{\underline{h}}$ and $G_{\underline{h}}$. Throughout this section, let q denote an arbitrary but fixed integer. Our main goal in this is to estimate quadratic sums corresponding to a general system of quadratic polynomials F, G defined as

$$T(q, \underline{z}) := \sum_{\underline{a}}^{q\,*} \sum_{\underline{y} \in \mathbb{Z}^n} \omega(\underline{y}/P) e((a_1/q + z_1)F(\underline{y}) + (a_2/q + z_2)G(\underline{y})). \tag{8.1}$$

Here F and G denote a system of quadratic polynomials with integer coefficients and ω denotes a compactly supported function on $\mathbb{R}^n$. Let us denote their leading quadratic parts by $F^{(0)}$ and $G^{(0)}$ respectively. We further assume that the quadratic forms $F^{(0)}$ and $G^{(0)}$ are defined by integer matrices M_1 and M_2 respectively. We will later apply the estimates in this section by setting $F = F_{\underline{h}}$ and $G = G_{\underline{h}}$.

Given a (finite or infinite) prime p, by m_p we denote

$$m_p := \max\{s_p(F^{(0)}), s_p(G^{(0)}), s_p(F^{(0)}, G^{(0)})\} \tag{8.2}$$

where further, given a set of forms $F_1, ..., F_R$, $s_p(F_1, ..., F_R)$ denotes the dimension of singular locus of the projective complete intersection variety defined by the simul-

taneous zero locus of the forms $F_1, ..., F_R$. That is:

$$s_p(F_1, ..., F_R) := \dim\{\underline{x} \in \mathbb{P}^n_{\overline{\mathbb{F}}_p} \ : \ F_1(\underline{x}) = \cdots = F_R(\underline{x}) = 0, \\ \text{Rank}_p(\nabla F_1(\underline{x}), \cdots, \nabla F_R(\underline{x})) < 2\}.$$

When $n \geq 2$, given an integer q, we define $D(q)$ by

$$D(q) := \prod_{\substack{p|q \\ p \text{ prime}}} p^{m_p+1}. \tag{8.3}$$

On the other hand, when $n = 1$, we define $D(q)$ as

$$D(q) := (q, \text{Cont}(F^{(0)}), \text{Cont}(G^{(0)})), \tag{8.4}$$

where, given a polynomial F, $\text{Cont}(F)$ is the gcd of all its coefficients.

As is standard ([3], [11], [12] for example), we begin by applying Poisson summation to $T(q, \underline{z})$. This will allow us separate the sum over $\underline{a}$ and the integral over $\underline{z}$, into an exponential sum and an exponential integral respectively. In particular, applying Poisson summation gives us the following:

Lemma 8.1. *We have*

$$T(q, \underline{z}) = q^{-n} \sum_{\underline{m} \in \mathbb{Z}} S(q; \underline{m}) I(\underline{z}; q^{-1}\underline{m})$$

where

$$S(q; \underline{m}, F, G) = S(q; \underline{m}) := \sum_{\underline{a}}^{q\,*} \sum_{\underline{u} \bmod q} e_q(a_1 F(\underline{u}) + a_2 G(\underline{u}) + \underline{m} \cdot \underline{u}), \tag{8.5}$$

and

$$I(\underline{\gamma}; \underline{k}) := \int_{\mathbb{R}^n} \omega(\underline{x}/P) e(\gamma_1 F(\underline{x}) + \gamma_2 G(\underline{x}) - \underline{k} \cdot \underline{x}) \, d\underline{x}. \tag{8.6}$$

Proof. The proof of Lemma 8.1 is standard and can be obtained by slightly modifying [3, Lemma 8]: Let $\underline{x} = \underline{u} + q\underline{v}$. Then

$$T(q, \underline{z}) = \sum_{\underline{a}}^{q\,*} \sum_{\underline{u} \bmod q} \sum_{\underline{v} \in \mathbb{Z}^n} \omega((\underline{u} + q\underline{v})/P) \times \\ e([a_1/q + z_1]F(\underline{u} + q\underline{v}) + [a_2/q + z_2]G(\underline{u} + q\underline{v}))$$

$$= \sideset{}{^*}\sum_{\underline{a}}^{q} \sum_{\underline{u} \bmod q} e_q(a_1F(\underline{u}) + a_2G(\underline{u})) \times$$

$$\sum_{\underline{v}\in\mathbb{Z}^n} \omega((\underline{u}+q\underline{v})/P)e(z_1F(\underline{u}+q\underline{v}) + z_2G(\underline{u}+q\underline{v})).$$

We now apply Poisson summation on the second sum (and use the substitution $\underline{x} = \underline{u} + q\underline{v}$) to get

$$\begin{aligned} T(q,\underline{z}) &= \sideset{}{^*}\sum_{\underline{a}}^{q} \sum_{\underline{u} \bmod q} e_q(a_1F(\underline{u}) + a_2G(\underline{u}))\times \\ &\qquad \sum_{\underline{m}\in\mathbb{Z}^n} \int_{\mathbb{R}^n} \omega((\underline{u}+q\underline{v})/P)e(z_1F(\underline{u}+q\underline{v}) + z_2G(\underline{u}+q\underline{v}) - \underline{m}\cdot\underline{v})\, d\underline{v} \\ &= q^{-n} \sum_{\underline{m}\in\mathbb{Z}^n} \sideset{}{^*}\sum_{\underline{a}}^{q} \sum_{\underline{u} \bmod q} e_q(a_1F(\underline{u}) + a_2G(\underline{u}) + \underline{m}\cdot\underline{u})\times \\ &\qquad \int_{\mathbb{R}^n} \omega(\underline{x}/P)e(z_1F(\underline{x}) + z_2G(\underline{x}) - q^{-1}\underline{m}\cdot\underline{x})\, d\underline{x} \end{aligned}$$

as required. □

As a result, we trivially have the following pointwise bound

$$|T(q,\underline{z})| \le q^{-n} \sum_{\underline{m}\in\mathbb{Z}} |S(q;\underline{m})| \cdot |I(\underline{z};q^{-1}\underline{m})|. \tag{8.7}$$

The treatment of the exponential integral is standard. In particular, we can use the following lemma to bound $I(\underline{z};q^{-1}\underline{m})$:

Lemma 8.2. *Let F, G be quadratic polynomials such that $\max\{\|F\|_P, \|G\|_P\} \ll H$, where*

$$\|F\|_P := \|P^{-\deg(F)}F(Px_1,\cdots,Px_n)\|. \tag{8.8}$$

Let $V := 1 + qP^{\varepsilon-1}\max\{1, HP^2|\underline{z}|\}^{1/2}$, $\varepsilon > 0$, and $N \in \mathbb{N}$. Then

$$I(\underline{z};q^{-1}\underline{m}) \ll_N P^{-N} + \mathrm{meas}(\{\underline{y} \in P\,\mathrm{Supp}(\omega_{\underline{h}}) :\ |\nabla\hat{F}_{\underline{z}}(\underline{y}) - \underline{m}| \le V\}),$$

where

$$\hat{F}_{\underline{z}}(\underline{x}) := qP^{-1}z_1F(\underline{x}) + qP^{-1}z_2G(\underline{x}).$$

Furthermore, if $|\underline{m}| \geq qP^{\epsilon-1}\max\{1, HP^2|\underline{z}|\}$*, then we have*

$$I(\underline{z}; q^{-1}\underline{m}) \ll_N P^{-N}|\underline{m}|^{-N}.$$

The proof of this is almost identical to the proofs of [2, Lemma 6.5-6.6], and so we will not detail it here. In particular, the only thing in the proofs that needs to be tweaked in order to verify Lemma 8.2 is that Θ in their equation (6.11) must be replaced with

$$\Theta' := 1 + |z_1|HP^2 + |z_2|HP^2.$$

We also note that we use $|\nabla \hat{F}_{\underline{z}}(\underline{y}) - \underline{m}| \leq V$ instead of $Pq^{-1}|\nabla \hat{F}_{\underline{z}}(\underline{y}) - \underline{m}| \leq Pq^{-1}V$ since we are using slightly different notation.

The latter bound enables us to handle the tail of the sum over $\underline{m}$. Let $\hat{V} := qP^{\epsilon-1}\max\{1, HP^2|\underline{z}|\}$. By trivially bounding $|S(q;\underline{m})|$ by q^n, and setting $N \geq n+2$, it is easy to show that

$$q^{-n}\sum_{|\underline{m}|\gg\hat{V}} |S(q;\underline{m})| \cdot |I(\underline{z}; q^{-1}\underline{m})| \ll 1,$$

by the second half of Lemma 8.2. Hence,

$$\implies |T_{\underline{h}}(q,\underline{z})| \ll 1 + q^{-n}\sum_{|\underline{m}|\ll\hat{V}} |S(q;\underline{m})| \cdot |I(\underline{z}; q^{-1}\underline{m})|.$$

Now by the first half of Lemma 8.2 (setting $N \geq n+4$), we have

$$\begin{aligned}|T_{\underline{h}}(q,\underline{z})| &\ll 1 + q^{-n}\sum_{|\underline{m}|\ll\hat{V}} |S(q;\underline{m})| \cdot \mathrm{meas}(\{\underline{y} \in P\,\mathrm{Supp}(\omega) :\ |\nabla \hat{F}_{\underline{z}}(\underline{y}) - \underline{m}| \leq V\} \\ &= 1 + q^{-n}\sum_{|\underline{m}|\ll\hat{V}} |S(q;\underline{m})| \int_{\underline{y}\in P\,\mathrm{Supp}(\omega)} \mathrm{Char}_G(\underline{m},\underline{y})\, d\underline{y},\end{aligned}$$

where

$$\mathrm{Char}_G(\underline{m},\underline{y}) = \begin{cases} 1 & \text{if } |\nabla \hat{F}_{\underline{z}}(\underline{y}) - \underline{m}| \leq V \\ 0 & \text{else.}\end{cases}$$

$$\implies |T_{\underline{h}}(q,\underline{z})| \ll 1 + q^{-n}\int_{\underline{y}\in P\,\mathrm{Supp}(\omega)} \sum_{\substack{|\underline{m}|\ll\hat{V} \\ |\nabla \hat{F}_{\underline{z}}(\underline{y}) - \underline{m}| \leq V}} |S(q;\underline{m})|\, d\underline{y}$$

$$\ll 1 + q^{-n} \int_{\underline{y} \in P\,\mathrm{Supp}(\omega)} \sum_{|\underline{m} - \underline{m}_0(\underline{y})| \leq V} |S(q; \underline{m})| \, d\underline{y}.$$

where $\underline{m}_0(\underline{y}) := \nabla \hat{F}_{\underline{z}}(\underline{y})$. Hence, we have the following:

Proposition 8.3. *Let* $|\underline{z}| = \max\{|z_1|, |z_2|\}$. *Then for any* $q \in \mathbb{N}$,

$$|T(q, \underline{z})| \ll 1 + q^{-n} \max_{\underline{y} \in P\,\mathrm{Supp}(\omega)} \Big\{ \sum_{|\underline{m} - \underline{m}_0(\underline{y})| \leq V} |S(q; \underline{m})| \Big\}.$$

for some $\underline{m}_0(\underline{y})$, *where*

$$V := 1 + qP^{-1+\varepsilon} \max\{1, HP^2|\underline{z}|\}^{1/2}.$$

Our attention now turns to finding a suitable bound for $|S(q; \underline{m})|$. As is standard when dealing with exponential sum bounds, we will take advantage of the multiplicative property of $S(q; \underline{m})$ and decompose q into its square-free, square, and cube-full components so that we can use better bounds in the former two cases (in particular, we will make use of the $\underline{a}$ sum to improve our bounds in the former cases). Indeed, we may use a Lemma of Hooley [15, Lemma 3.2] to get the following result:

Lemma 8.4. *Let* $\underline{a} \in \mathbb{Z}^2$ *s.t.* $(q, \underline{a}) = 1$, $q = rs$ *where* $(r, s) = 1$ *and* $\underline{m} \in \mathbb{Z}^n$. *Then*

$$S(rs; \underline{m}) = S(r; \bar{s}\underline{m}) S(s; \bar{r}\underline{m}), \tag{8.9}$$

where $r\bar{r} + s\bar{s} = 1$.

Proof. The claim is trivial when $r = q$ or $s = q$, so we will assume $r, s \neq q$. Before we proceed with the proof, we will firstly show that $(r\bar{r})^j \equiv r\bar{r} \mod q$, $(s\bar{s})^j \equiv s\bar{s} \mod q$ for every $j \in \mathbb{N}$. Indeed, since $r\bar{r} + s\bar{s} = 1$, we automatically have that $(r\bar{r})^j \equiv r\bar{r} \equiv 1 \mod s$. Hence $r(r\bar{r})^j \equiv r(r\bar{r}) \mod q$ by $q = rs$, and since $r \neq q$, this gives us $(r\bar{r})^j \equiv r\bar{r} \mod q$ for $j \in \mathbb{N}$.

Now if we let $\underline{x} = r\bar{r}\underline{s} + s\bar{s}\underline{r}$, where $\underline{r} \in (\mathbb{Z}/r\mathbb{Z})^n, \underline{s} \in (\mathbb{Z}/s\mathbb{Z})^n$, then $(r\bar{r})^j \equiv r\bar{r} \mod q$, $(s\bar{s})^j \equiv s\bar{s} \mod q$, and $r\bar{r} + s\bar{s} = 1$ implies that

$$a_1 F(\underline{x}) + a_2 G(\underline{x}) \equiv r\bar{r}(a_1 F(\underline{s}) + a_2 G(\underline{s}) + \underline{m} \cdot \underline{s}) +$$

$$s\bar{s}(a_1F(\underline{r}) + a_2G(\underline{r}) + \underline{m}\cdot\underline{r}) \mod q.$$

In particular, we have

$$e_q(a_1F(\underline{x}) + a_2G(\underline{x}) + \underline{m}\cdot\underline{x}) = e_s(\bar{r}(a_1F(\underline{s})+a_2G(\underline{s}) + \underline{m}\cdot\underline{s}))\times$$
$$e_r(\bar{s}(a_1F(\underline{r}) + a_2G(\underline{r}) + \underline{m}\cdot\underline{r})).$$

Hence

$$S(rs;\underline{m}) = \sum_{\underline{a}}^{q\,*} \sum_{\underline{s} \bmod s} \sum_{\underline{r} \bmod r} e_s(\bar{r}(a_1F(\underline{s})+a_2G(\underline{s}) + \underline{m}\cdot\underline{s}))\times \tag{8.10}$$
$$e_r(\bar{s}(a_1F(\underline{r}) + a_2G(\underline{r}) + \underline{m}\cdot\underline{r})).$$

Finally, we may set $\underline{a} = r\underline{u}+s\underline{v}$, $\underline{u} \in (Z/s\mathbb{Z})^2$, $\underline{v} \in (\mathbb{Z}/r\mathbb{Z})^2$, and note that $(u_1, u_2, s) = (v_1, v_2, r) = 1$ since $(a_1, a_2, q) = 1$ is true if and only if $(a_1, a_2, r) = (a_1, a_2, s) = 1$ (recall that r, s are coprime). Hence by (8.10) and recalling that $r\bar{r} \equiv 1 \mod s$, $s\bar{s} \equiv 1 \mod r$:

$$\begin{aligned} S(rs;\underline{m}) &= \sum_{\underline{u}}^{*} \sum_{\underline{v}}^{*} \sum_{\underline{s} \bmod s} \sum_{\underline{r} \bmod r} e_s(r\bar{r}(u_1F(\underline{s}) + u_2G(\underline{s})) + \bar{r}\underline{m}\cdot\underline{s})\times \\ &\qquad e_r(s\bar{s}(v_1F(\underline{r}) + v_2G(\underline{r})) + \bar{s}\underline{m}\cdot\underline{r}) \\ &= \Big(\sum_{\underline{u}}^{*} \sum_{\underline{s} \bmod s} e_s(u_1F(\underline{s}) + u_2G(\underline{s}) + \bar{r}\underline{m}\cdot\underline{s})\Big)\times \\ &\qquad \Big(\sum_{\underline{v}}^{*} \sum_{\underline{r} \bmod r} e_r(v_1F(\underline{r}) + v_2G(\underline{r}) + \bar{s}\underline{m}\cdot\underline{r})\Big) \\ &= S(r;\bar{s}\underline{m})S(s;\bar{r}\underline{m}) \end{aligned}$$

as required. □

Our treatment of bounds for the quadratic exponential sums will vary depending on whether q is square-free, a square or cube-full. Since the exponential sums satisfy

the mutliplicativity relation (8.9), it is natural to set $q = b_1 b_2 q_3$ where

$$b_1 := \prod_{p||q} p, \quad b_2 := \prod_{p^2||q} p^2, \quad q_3 := \prod_{\substack{p^e||q \\ e>2}} p^e. \tag{8.11}$$

Then by Lemma 8.4, we have that

$$S(q; \underline{m}) = S(b_1; c_1\underline{m})S(b_2; c_2\underline{m})S(q_3; c_3\underline{m}), \tag{8.12}$$

for some constants c_1, c_2, c_3 such that $(b_1, c_1) = (b_2, c_2) = (q_3, c_3) = 1$. Finding suitable bounds for the size of these three exponential sums will be the topic of the rest of this section.

8.1 Square-free Exponential Sums

In this subsection, we will briefly consider the quadratic exponential sums $S(b_1; \underline{m})$ when $q = b_1$ is square-free. This case is extensively studied in [21, Section 5], where bounds are obtained for exponential sums for a general system of polynomials F and G. Using the multiplicativity of the exponential sum in (8.9), it is enough to consider the sums $S(p, \underline{m})$ where p is a prime. We may rewrite

$$S(p, \underline{m}) = \Sigma_1 - \Sigma_4, \tag{8.13}$$

where

$$\Sigma_1 := \sum_{a_1=1}^{p} \sum_{a_2=1}^{p} \sum_{\underline{u} \bmod q} e_p(a_1 F(\underline{u}) + a_2 G(\underline{u}) + \underline{m} \cdot \underline{u}) \quad \text{and} \quad \Sigma_4 := \sum_{\underline{u} \bmod q} e_p(\underline{m} \cdot \underline{u}). \tag{8.14}$$

Here the notation Σ_1 and Σ_4 is used to correspond to the corresponding sums in [21, Section 5]. Note that the argument in [21, Section 5] does not depend on the degree of the forms F and G. In fact our exponential sums are more "natural" than the ones which appear in [21] and as a result, only sums Σ_1 and Σ_4 appear in our analysis. We may now use the results in [21, Section 5] directly here as they do indeed bound the sums Σ_1 and Σ_4 as well, but only in the case where F and G

intersect properly over $\overline{\mathbb{F}}_p$. When $n \geq 2$, we may use [21, Prop 5.2, Lemma 5.4] to get

Proposition 8.5. *Let $F, G \in \mathbb{Z}[x_1, \cdots, x_n]$ be quadratic polynomials such that $m_\infty(F^{(0)}, G^{(0)}) = -1$. Let b_1 be a square-free number where*

$$(b_1, \mathrm{Cont}(F^{(0)})) = (b_1, \mathrm{Cont}(G^{(0)})) = 1,$$

If $n > 1$, then there exists some $\Phi_{F,G} = \Phi \in \mathbb{Z}[x_1, \cdots, x_n]$ such that

$$S(b_1, \underline{m}) \ll_n b_1^{1+n/2+\varepsilon} D(b_1)(b_1, \Phi(\underline{m}))^{1/2}$$

for every $\underline{m} \in \mathbb{Z}^n$. Furthermore Φ has the following properties:

1. *Φ is homogeneous.*
2. *$\deg(\Phi) \ll_n 1$.*
3. *$\log ||\Phi|| \ll_n \log ||F|| + \log ||G||$.*
4. *$\mathrm{Cont}(\Phi) = 1$.*

In the quadratic case at hand, the dual variety Φ could be made more explicit, but, it is not required in this work.

Proof. In the case when F, G intersect properly over $\overline{\mathbb{F}}_p$, [21, Prop 5.2, Lemma 5.4] hands us

$$S(p, \underline{m}) \ll_n p^{1+n/2+\varepsilon} D(p)^{1/2} (p, \Phi(\underline{m}))^{1/2}. \tag{8.15}$$

We note that we have $D(b_1)^{1/2}$ appearing, whilst in [21], $D(b_1)$ appears instead. This is due to Marmon and Vishe using a different definition for $D(b_1)$, namely

$$D(b_1) := \prod_{\substack{p | b_1 \\ p \text{ prime}}} p^{(m_p+1)/2},$$

whilst in our case, we use p^{m_p+1} instead. In our context, it is more natural to define $D(b_1)$ as we do in (8.3). This is because – unlike in [21] – both of our forms F, G vary

as $\underline{h}$ varies, and this forces us to consider the case when F and G intersect improperly in more detail. In particular, the bound we find is more naturally expressed by using (8.3) as our definition of $D(b_1)$.

In the case where $n > 1$ and F, G intersect improperly over $\overline{\mathbb{F}}_p$, we either have $m_p(F,G) = n-2$ or $m_p(F,G) = n-1$ by Lemma 4.1. When $m_p(F,G) = n-1$, we must have that at least one of F, G are equal to the zero polynomial over $\mathbb{Z}/p\mathbb{Z}$. We bound trivially in this case:

$$\begin{aligned}|S(p;\underline{m})| &\leq p^{n+2}\\ &= p^{1+n/2}p^{1+n/2}\\ &\leq p^{1+n/2}p^{(n-1)+1}\\ &= p^{1+n/2}p^{m_p+1} = p^{1+n/2}D(p),\end{aligned}$$

since $n \geq 2$. Hence the only thing left to consider is the case where $m_p(F,G) = n-2$. In this case, $F, G \not\equiv 0 \mod p$, and so we certainly have

$$\Sigma_1 \ll p^2 \sum_{\substack{\underline{x}\\ p|F(\underline{x})}}^{p} 1 \leq p^{n+1} \leq p^{1+n/2}D(p),$$

again since $n \geq 2$. Finally, we recall (8.13) and note that $|\Sigma_4| \leq p^n$. Hence

$$|S(p;\underline{m})| \ll p^{1+n/2}D(p). \tag{8.16}$$

Therefore, we may conclude that for a general p (irrespective of whether or not the intersection is proper)

$$S(p,\underline{m}) \leq C(n)p^{1+n/2+\varepsilon}D(p)(p,\Phi(\underline{m}))^{1/2},$$

where C is some constant. Finally by Lemma 8.4, we have

$$\begin{aligned}S(b_1,\underline{m}) &= \prod_{p|b_1} S(p,c_p\underline{m})\\ &\leq C(n)^{d(b_1)}b_1^{1+n/2+\varepsilon}D(b_1)\prod_{p|b_1}(p,\Phi(c_p\underline{m}))^{1/2}\\ &= C(n)^{d(b_1)}b_1^{1+n/2+\varepsilon}D(b_1)(b_1,\Phi(\underline{m}))^{1/2},\end{aligned}$$

where $d(b_1) := \#\{p \mid b_1\}$ is the divisor function of b_1. We could replace $(p, \Phi(c_p\underline{m}))$ with $(b_1, \Phi(\underline{m}))$ because Φ is homogeneous and $(p, c_p) = 1$. All that is left to do is show that $C(n)^{d(b_1)}$ does not contribute more than $O(P^\varepsilon)$. To see this, we note that $d(b_1) \ll \log(b_1)/\log\log(b_1)$. Hence there is some constant d such that

$$\begin{aligned} C(n)^{d(b_1)} &\leq C(n)^{d\log(b_1)/\log\log(b_1)} \\ &= e^{\log(C^{[d\log(b_1)/\log\log(b_1)]})} \\ &= e^{d\log(b_1)\log(C)/\log\log(b_1)} \\ &= e^{\log(b_1^{[d\log(C)/\log\log(b_1)]})} \\ &= b_1^{d\log(C)/\log\log(b_1)} \ll b_1^\varepsilon \end{aligned}$$

provided that $b_1 \gg_\varepsilon 1$. We automatically have $d(b_1) \ll 1$ if $b_1 \not\gg 1$, so we get $c^{d(b_1)} \ll 1 \ll b_1^\varepsilon$ in that case. Hence, we may conclude that Proposition 8.5 is true. We will bound the $C(n)$ term in future lemmas by b_i^ϵ without further comment. □

We also must consider when $n = 1$. In this case, it is sufficient for us to use a weaker bound than [21, Lemma 5.5]. We will show the following:

Proposition 8.6. *Let $F, G \in \mathbb{Z}[x]$ be quadratic polynomials and let b_1 be a square-free integer. Then*

$$S(b_1, m) \ll b_1^{2+\varepsilon} D(b_1).$$

Proof. The proof of Proposition 8.6 is almost trivial. We start by applying Lemma 8.4 so that we may consider $S(p; cm)$ for some $p \nmid c$. We note that

$$|\Sigma_1| = p^2 \#\{x \mod p \,:\, F(x) \equiv G(x) \equiv 0 \mod p\} \ll p^2(p, \mathrm{Cont}(F), \mathrm{Cont}(G)),$$

and we trivially have $|\Sigma_4| \leq p$. Hence, by (8.4) and noting that $(p, \mathrm{Cont}(F), \mathrm{Cont}(G)) \leq (p, \mathrm{Cont}(F^{(0)}), \mathrm{Cont}(G^{(0)}))$:

$$|S(p; cm)| \leq |\Sigma_1| + |\Sigma_4| \ll p^2 D(p),$$

and so

$$|S(b_1; m)| \ll b_1^{2+\varepsilon} D(b_1)$$

for any $m \in \mathbb{Z}$. □

8.2 Square-full Bound

In this section, we will derive the bound which will be used when q is square-full. When q is square-full, we give up on saving q over the $\underline{a}$ sum, and start with the bound

$$|S(q;\underline{m})| \leq \sum_{\underline{a}}^{q*} |S(\underline{a},q;\underline{m})| \tag{8.17}$$

where F, G are quadric polynomials, and

$$S(\underline{a},q;\underline{m}) := \sum_{\underline{x} \bmod q} e_q(a_1F(\underline{x}) + a_2G(\underline{x}) + \underline{m}\cdot\underline{x}).$$

For a fixed value of $\underline{a}$, the exponential sum $S(\underline{a},q;\underline{m})$ is a standard quadratic exponential sum with leading quadratic part defined by the matrix

$$M(\underline{a}) := M := a_1M_1 + a_2M_2, \tag{8.18}$$

as defined in (4.8). We will assume further that $2 \,|\, (\mathrm{Cont}(F^{(0)}), \mathrm{Cont}(G^{(0)}))$ so that $M(\underline{a}) \in M_n(\mathbb{Z})$ for every $\underline{a}$.

Remark 8.7. *In the broader context of the argument that we are building, the reason why we may assume that* $2 \,|\, (\mathrm{Cont}(F^{(0)}), \mathrm{Cont}(G^{(0)}))$ *is due to Remark 5.1: If the coefficients of our original cubic forms in Chapter 5 are divisible by* 2*, then the coefficients of the differenced quadratic polynomials coming from Chapter 7 must also be divisible by* 2*.*

A standard squaring argument as obtained in [28, Lemma 2.5] for example readily hands us a bound

$$|S(\underline{a},q;\underline{m})| \ll q^{n/2}\#\mathrm{Null}_q(M)^{1/2}, \tag{8.19}$$

where $\#\mathrm{Null}_q(M)$ denotes the number of solutions of the equation $M\underline{x} \equiv \underline{0} \bmod q$ as defined in (4.16). To estimate this, we will resort to using a Smith normal form

of the matrix M. The Smith normal form of M hands us invertible integer matrices S and T be with determinant ± 1 such that

$$SMT = \mathrm{Smith}(M) = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & \cdots & & \lambda_n \end{pmatrix} \in M_n(\mathbb{Z}), \tag{8.20}$$

where $\lambda_1 \mid \lambda_2 \mid \cdots \mid \lambda_n$. Since the forms $F^{(0)}$ and $G^{(0)}$ are assumed to be arbitrary for now, it is easy to conclude that

$$|S(\underline{a}, q; \underline{m})| \ll q^{n/2} \prod_{i=1}^{n} \lambda_{q,i}^{1/2}, \tag{8.21}$$

where

$$\lambda_{q,i} := (q, \lambda_i). \tag{8.22}$$

Remark 8.8. *Recall that we aim to finally substitute $F = F_{\underline{h}}$ and $G = G_{\underline{h}}$. Note that the extra factor appearing on the right hand side of* (8.21) *is a generalisation of the factor $D(b_1)^{1/2}$ appearing in Proposition 8.5. This is a drawback of van der Corput differencing that although one starts with a nice pair of forms F and G, one ends up with exponential sums of differenced polynomials $F_{\underline{h}}$ and $G_{\underline{h}}$, which can be highly singular modulo q. If $q = p^\ell$ for some prime p, if the singular locus m_p as defined in* (8.2) *is large, then this gives restrictions on the vector $\underline{h} \bmod p$. When ℓ is small, the extra factors appearing can be compensated from the corresponding bounds on the $\underline{h}$ sum. However, in the case at hand, when $q = p^\ell$ for a large ℓ, we can not rule out the possibility that for many $\underline{h}$, there may exist a large q such that the factor $\prod_{i=1}^{n} \lambda_{q,i}^{1/2}$ is as large as $q^{n/2}$. This complication arises partly due to the simplicity of the quadratic exponential sums appearing. However, later we would need to average the sums over various $|\underline{m} - \underline{m}_0| \leq V$. We will aim to salvage some of this loss by gaining a congruence condition on $\underline{m}$ instead and saving from the sum over $\underline{m}$. This idea partly has already featured in Vishe's work [28, Lemma 6.4]. However, in [28],*

the author is dealing with fixed F and G, which is not the case here.

Our main goal here is to prove the following result:

Proposition 8.9. *Let $\underline{a} \in \mathbb{Z}^2$ and $q \in N$ be such that $(\underline{a}, q) = 1$, let $\underline{m} \in \mathbb{Z}^n$, and let F, G be quadratic polynomials such that $2 \mid (\mathrm{Cont}(F^{(0)}), \mathrm{Cont}(G^{(0)}))$ (see Remark 8.7). Let*

$$(a_1F_1 + a_2F_2)(\underline{x}) = \underline{x}^t M \underline{x} + \underline{\mathfrak{b}} \cdot \underline{x} + \mathfrak{c}, \tag{8.23}$$

(We use $\underline{\mathfrak{b}}$ instead of $\underline{b}$ to avoid confusion since we have already defined b_1, b_2, b_3). Then

$$|S(\underline{a}, q; \underline{m})| \leq 2^{n/2} q^{n/2} \#\mathrm{Null}_q(M)^{1/2} \Delta_q(\underline{m} + \underline{\mathfrak{b}})$$

where

$$\Delta_q(\underline{m}) := \Delta_{T,q}(\underline{m}) := \begin{cases} 1 & \text{if } \lambda_{q,i} \mid (T^t \underline{m})_i \text{ for } 1 \leq i \leq n \\ 0 & \text{else.} \end{cases} \tag{8.24}$$

Here, T be the matrix appearing in the Smith normal form of M in (8.20), $\lambda_{q,i}$ be as in (8.22) and given a vector $\underline{v}$, let $(\underline{v})_i$ denote its i-th component.

Proof. To estimate $|S(\underline{a}, q; \underline{m})|$, we begin by working with its square:

$$\begin{aligned} |S(\underline{a}, q, \underline{m})|^2 &= \sum_{\underline{x}, \underline{y} \bmod q} e_q((a_1F_1 + a_2F_2)(\underline{x}) + \underline{m} \cdot \underline{x}) \overline{e_q((a_1F_1 + a_2F_2)(\underline{y}) + \underline{m} \cdot \underline{y})} \\ &= \sum_{\underline{x}, \underline{y} \bmod q} e_q(\underline{x}^t M \underline{x} - \underline{y}^t M \underline{y} + (\underline{m} + \underline{\mathfrak{b}}) \cdot (\underline{x} - \underline{y})). \end{aligned}$$

We will now change order of summation by setting $\underline{x} = \underline{y} + \underline{z}$. Then

$$\begin{aligned} |S(\underline{a}, q, \underline{m})|^2 &= \sum_{\underline{y}, \underline{z} \bmod q} e_q(\underline{z}^t M \underline{z} + (\underline{m} + \underline{\mathfrak{b}}) \cdot \underline{z} + 2\underline{y}^t M \underline{z}) \\ &= \sum_{\underline{z} \bmod q} e_q(\underline{z}^t M \underline{z} + \underline{m}' \cdot \underline{z}) \sum_{\underline{y} \bmod q} e_q(\underline{y} \cdot 2M\underline{z}). \end{aligned}$$

where $\underline{m}' = \underline{m} + \underline{\mathfrak{b}}$. Therefore

$$|S(\underline{a}, q, \underline{m})|^2 = q^n \sum_{\underline{z} \bmod q} e_q(\underline{z}^t M \underline{z} + \underline{m}' \cdot \underline{z}) \delta_{2M}(\underline{z}), \tag{8.25}$$

where

$$\delta_M(\underline{z}) := \begin{cases} 1 & \text{if } M\underline{z} \equiv 0 \mod q \\ 0 & \text{else.} \end{cases} \tag{8.26}$$

The "2" appearing in $\delta_{2M}(\underline{z})$ gives rise to some minor technical difficulties in the case when q is even. Therefore, we will start by considering the case when q is odd first.

8.2.1 Case: q odd

In this case, $\delta_{2M}(\underline{z}) = 1$ if and only if $M\underline{z} \equiv \underline{0} \mod q$, and so we may replace $\delta_{2M}(\underline{z})$ in (8.25) by $\delta_M(\underline{z})$. Furthermore, we note that $M\underline{z} \equiv \underline{0} \mod q$ implies that $\underline{z}^t M\underline{z} \equiv \underline{0} \mod q$. Hence (8.25) simplifies as:

$$|S(\underline{a}, q, \underline{m})|^2 = q^n \sum_{\underline{z} \bmod q} e_q(\underline{m}' \cdot \underline{z})\delta_M(\underline{z}). \tag{8.27}$$

Now, M has a Smith Normal form over $\mathbb{Z}$ as in (8.20), $\text{Smith}(M) := SMT$, for some matrices $S, T \in SL_n(\mathbb{Z})$. In particular, matrices S and T are invertible over $\mathbb{Z}/q\mathbb{Z}$, for any $q \in \mathbb{N}$. We will now rewrite our sum in terms of the $\text{Smith}(M)$, Firstly, we note that

$$\delta_M = \delta_{SM}.$$

Therefore, on using the substitution $\underline{z} \mapsto T^{-1}\underline{z}$, (8.25) becomes

$$|S(\underline{a}, q, \underline{m})|^2 = q^n \sum_{\underline{z} \bmod q} e_q(\underline{m}' \cdot T\underline{z})\delta_{SMT}(\underline{z}), \tag{8.28}$$

since $\delta_{SM}(T\underline{z}) = \delta_{SMT}(\underline{z})$ by (8.26). We will now work towards determining which $\underline{z}$ make $\delta_{SMT}(\underline{z})$ non-zero. By definition, $\delta_{SMT}(\underline{z}) \neq 0$ if and only if

$$SMT\underline{z} \equiv \underline{0} \mod q,$$

or equivalently

$$\underline{z} \in \text{Null}_q(SMT) := \{\underline{x} \in \left(\mathbb{Z}/q\mathbb{Z}\right)^n \mid SMT\underline{x} \equiv \underline{0} \mod q\}.$$

Hence, we may simplify (8.28) as follows:

$$\begin{aligned} |S(\underline{a}, q, \underline{m})|^2 &= q^n \sum_{\underline{z} \in \mathrm{Null}_q(SMT)} e_q(\underline{m}' \cdot T\underline{z}) \\ &= q^n \sum_{\underline{z} \in \mathrm{Null}_q(SMT)} e_q(\underline{z} \cdot T^t \underline{m}') \end{aligned} \tag{8.29}$$

where T^t is the transpose of T. This is true because

$$\underline{m}' \cdot T\underline{z} = (T\underline{z})^t \underline{m}' = \underline{z}^t T^t \underline{m}' = \underline{z} \cdot T^t \underline{m}'.$$

We now turn our attention to structure of the $\mathrm{Null}_q(SMT)$. Since S and T are defined to be the unique matrices (up to units) such that $SMT = \mathrm{Smith}(M)$, it is quite easy to determine precisely when $\underline{z} \in \mathrm{Null}_q(SMT)$. Therefore $SMT\underline{z} \equiv \underline{0}$ mod q if and only if

$$\frac{q}{\lambda_{q,i}} \,\Big|\, z_i \tag{8.30}$$

for every $i \in \{1, \cdots, n\}$. Therefore

$$\#\mathrm{Null}_q(SMT) = \prod_{i=1}^{n} \lambda_{q,i}. \tag{8.31}$$

Hence by (8.22), and (8.29)-(8.30), we have the following:

$$\begin{aligned} |S(\underline{a}, q, \underline{m})|^2 &= q^n \prod_{i=1}^{n} \sum_{q/\lambda_{q,i} | z_i} e_q(z_i (T^t \underline{m}')_i) \\ &= q^n \prod_{i=1}^{n} \sum_{x_i=1}^{\lambda_{q,i}} e_{\lambda_{q,i}}(x_i (T^t \underline{m}')_i) \\ &= q^n \prod_{i=1}^{n} \lambda_{q,i} \delta_{q,i}(\underline{m}'), \end{aligned} \tag{8.32}$$

where

$$\delta_{q,i}(\underline{u}) := \begin{cases} 1 & \text{if } \lambda_{q,i} \mid (T^t \underline{u})_i \\ 0 & \text{else} \end{cases}, \tag{8.33}$$

and $(\underline{v})_i$ is the i-th component of vector $\underline{v}$. Therefore, by (8.31) and (8.32):

$$|S(\underline{a}, q, \underline{m})|^2 = q^n \#\mathrm{Null}_q(SMT) \prod_{i=1}^{n} \delta_{q,i}(\underline{m}').$$

Finally it is easy to check that

$$\#\mathrm{Null}_q(SMT) = \#\mathrm{Null}_q(M)$$

since S and T are both invertible over $\mathbb{Z}/q\mathbb{Z}$ and therefore in this case we establish:

$$|S(\underline{a}, q; \underline{m})| = q^{n/2} \#\mathrm{Null}_q(M)^{1/2} \Delta_q(\underline{m} + \underline{\mathfrak{b}}),$$

which clearly suffices.

8.2.2 Case: q even

We now turn to the case where q is even. In this case, the above argument needs to be modified due to not being able to directly replace the condition $\delta_{2M}(\underline{z})$ with $\delta_M(\underline{z})$ in (8.25). Instead we note that $\delta_{2M}(\underline{z}) \neq 0$ if and only if $M\underline{z} \equiv \underline{0} \mod q/2$. In particular, there must be some $\underline{c} \in \{0,1\}^n$ such that

$$M\underline{z} \equiv \frac{q}{2}\underline{c} \mod q.$$

Therefore, if we let

$$N_{\underline{c},q}(M) := \{\underline{x} \mod q \ : \ M\underline{x} \equiv \frac{q}{2}\underline{c} \mod q\},$$

then $\delta_{2M(\underline{z})} \neq 0$ if and only if $\underline{z} \in N_{\underline{c},q}$ for some $\underline{c}$. Hence, we may rewrite (8.25) as follows:

$$|S(\underline{a}, q; \underline{m})|^2 = q^n \sum_{\underline{c} \in \{0,1\}^n} \sum_{\underline{z} \in N_{\underline{c},q}(M)} e_q(\underline{z}^t M \underline{z} + \underline{m}' \cdot \underline{z}) \tag{8.34}$$

We now wish to write $N_{\underline{c},q}$ in terms of $\mathrm{Null}_q(M)$ as this will enable us to use the arguments discussed in the odd case. To do this, we invoke Lemma 4.7 to see that either $N_{\underline{c},q} = \emptyset$ or there exists some $\underline{y}_{\underline{c}} \in (\mathbb{Z}/q\mathbb{Z})^n$ such that

$$N_{\underline{c},q} = \underline{y}_{\underline{c}} + \mathrm{Null}_q(M).$$

Hence

$$\begin{aligned}
|S(\underline{a},q;\underline{m})|^2 &= q^n \sum_{\substack{\underline{c}\in\{0,1\}^n \\ N_{\underline{c},q}(M)\neq\emptyset}} \sum_{\underline{z}\in\underline{y}_{\underline{c}}+\mathrm{Null}_q(M)} e_q(\underline{z}^t M\underline{z}+\underline{m}'\cdot\underline{z}) \\
&= q^n \sum_{\substack{\underline{c}\in\{0,1\}^n \\ N_{\underline{c},q}(M)\neq\emptyset}} \sum_{\underline{z}\in\mathrm{Null}_q(M)} e_q([\underline{y}_{\underline{c}}+\underline{z}]^t M[\underline{y}_{\underline{c}}+\underline{z}]+\underline{m}'\cdot[\underline{y}_{\underline{c}}+\underline{z}]) \\
&= q^n \sum_{\substack{\underline{c}\in\{0,1\}^n \\ N_{\underline{c},q}(M)\neq\emptyset}} e_q(\underline{y}_{\underline{c}}^t M\underline{y}_{\underline{c}}+\underline{m}'\cdot\underline{y}_{\underline{c}}) \sum_{\underline{z}\in\mathrm{Null}_q(M)} e_q((\underline{z}+2\underline{y}_{\underline{c}})^t M\underline{z}+\underline{m}'\cdot\underline{z}) \\
&\leq q^n \sum_{\underline{c}\in\{0,1\}^n} \Big| \sum_{\underline{z}\in\mathrm{Null}_q(M)} e_q((\underline{z}+2\underline{y}_{\underline{c}})^t M\underline{z}+\underline{m}'\cdot\underline{z})\Big|. \qquad (8.35)
\end{aligned}$$

Finally, we note that $M\underline{z}\equiv\underline{0} \mod q$ since $\underline{z}\in\mathrm{Null}_q(M)$, and so by (8.35), we have the following:

$$\begin{aligned}
|S(\underline{a},q;\underline{m})|^2 &\leq q^n \sum_{\underline{c}\in\{0,1\}^n} \Big| \sum_{\underline{z}\in\mathrm{Null}_q(M)} e_q(\underline{m}'\cdot\underline{z})\Big| \\
&= 2^n q^n \Big| \sum_{\underline{z} \bmod q} e_q(\underline{m}'\cdot\underline{z})\delta_M(\underline{z})\Big|.
\end{aligned}$$

This is precisely (8.27) with an extra factor of 2^n and some absolute value signs around the sum (which are irrelevant). We may therefore repeat the arguments in the q odd case which follow from (8.27) to establish Proposition 8.9. □

8.2.3 Special Case: $n = 1$

We will now briefly consider the case when $n = 1$, as we will need to deal with this case separately later. The arguments used above are still valid in this case, but the bound that we get is simpler due to the matrix, M, becoming an integer. In particular, Proposition 8.9 becomes

Proposition 8.10. *Let $\underline{a}\in\mathbb{Z}^2$ and $q\in N$ be such that $(\underline{a},q)=1$, let $m\in\mathbb{Z}$, and let $F,G\in\mathbb{Z}[x]$ be quadratic polynomials. Let*

$$(a_1F_1+a_2F_2)(x) = Mx^2+bx+c. \qquad (8.36)$$

Then

$$|S(\underline{a}, q; \underline{m})| \leq 2^{1/2} q^{1/2} (q, M)^{1/2} \Delta_q'(m + b)$$

where

$$\Delta_q'(m) := \begin{cases} 1 & \text{if } (q, M) \mid m \\ 0 & \text{else.} \end{cases} \tag{8.37}$$

We will use Propositions 8.9 and 8.10 directly in our future treatment of the cube-full part of $S(q_3, \underline{m})$ (see (8.12)) in order to get additional saving over the $\underline{m}$ sum. For the *perfect square* part – b_2 – however, we will derive a slightly weaker bound from this which will be used to get saving over the $\underline{h}$ sum later on in the argument.

8.3 Cube-free Square Exponential Sums

In this subsection, we will assume that $q = b_2$, or equivalently q is a cube-free square. In this case, we will give up on the potential saving we could attain via the $\underline{m}$ sum from the $\Delta_q(\underline{m}')$ term in Proposition 8.9, and bound $\#\text{Null}_q(M(\underline{a}))^{1/2}$ in terms of the singular locus of F, G, where $M(\underline{a})$ is defined as in (8.23). In this special case, we will need to get pointwise saving over the $\underline{a}$ sum in order for our bound to be useful. We will start with the case when $n \geq 2$. Upon letting $b_2 = c^2$, and by Proposition 8.9, Lemmas 4.5 - 4.6, and (8.17) we have

$$\begin{aligned} |S(b_2, \underline{m})| \leq \sideset{}{^*}\sum_{\underline{a} \bmod b_2} |S(\underline{a}, b_2; \underline{m})| &\leq b_2^{n/2} \sideset{}{^*}\sum_{\underline{a} \bmod b_2} \#\text{Null}_{c^2}(M(\underline{a}))^{1/2} \\ &\leq b_2^{n/2} \sideset{}{^*}\sum_{\underline{a} \bmod b_2} \#\text{Null}_{c}(M(\underline{a})) \\ &\ll b_2^{2+n/2} c^{m_p+1} \\ &= b_2^{2+n/2} \prod_{\substack{p^2 \mid q \\ p \text{ prime}}} p^{m_p+1} \\ &= b_2^{2+n/2} D(b_2). \end{aligned} \tag{8.38}$$

When $n = 1$, we have $M(\underline{a}) = a_1 d_F + a_2 d_G$ for some constants d_F, d_G. the same type of argument applies. By Proposition 8.10,

$$\begin{aligned} |S(p^2, \underline{m})| &\leq p \sideset{}{^*}\sum_{\underline{a}}^{p^2} (p^2, M(\underline{a}))^{1/2} \\ &\leq p \sideset{}{^*}\sum_{\underline{a}}^{p^2} (p, a_1 d_F + a_2 d_G) \\ &= p\Big(\sideset{}{^*}\sum_{p \mid a_1 d_F + a_2 d_G}^{p^2} p + \sideset{}{^*}\sum_{p \nmid a_1 d_F + a_2 d_G}^{p^2} 1 \Big) \\ &\leq \begin{cases} 2p^5 \text{ if } (d_F, d_G, p) = 1 \\ p^6 \text{ else} \end{cases} \end{aligned} \tag{8.39}$$

Hence, upon recalling (8.4), we may bound (8.39) by

$$|S(p^2, \underline{m})| \ll p^5 D(p).$$

We may then use Lemma 8.4 and the argument from Proposition 8.6 to get

$$|S(b_2, \underline{m})| \ll b_2^{2+1/2+\varepsilon} D(b_2).$$

Combining this with (8.38) gives us the following:

Proposition 8.11. *Let $b_2 \in \mathbb{N}$ be a cube-free square. Then*

$$S(b_2, \underline{m}) \ll b_2^{2+n/2+\varepsilon} D(b_2).$$

Chapter 9

Quadratic Exponential Sums: Finalisation

In this section, we will combine all of the bounds we have found in Section 8 to reach our final estimate for $T(q, \underline{z})$. Recall that Proposition 8.3 hands us

$$|T(q,\underline{z})| \ll 1 + q^{-n} \max_{\underline{y} \in P\,\mathrm{Supp}(\omega)} \Big\{ \sum_{|\underline{m}-\underline{m}_0(\underline{y})|\leq V} |S(q;\underline{m})| \Big\}. \tag{9.1}$$

In the previous chapter, we focused on getting bounds for individual exponential sums $|S(q;\underline{m})|$. In this chapter, we will apply those bounds to (9.1) and then aim to attain saving over the $\underline{m}$ sum.

We begin by considering averages of exponential sums. Throughout, let $\underline{m}_0$ be an arbitrary but fixed vector in $\mathbb{Z}^n$ and let $\underline{\mathfrak{b}}(\underline{a}) = \underline{\mathfrak{b}}$ be defined as in (8.23). For $n \geq 2$: By Lemma 8.4 and Propositions 8.5, 8.9, and 8.11, there are some constants c_1, c_2, c_3 such that $(b_1, c_1) = (b_2, c_2) = (q_3, c_3) = 1$, and

$$\begin{aligned}\sum_{|\underline{m}-\underline{m}_0|\leq V} |S(q;\underline{m})| &\leq \sum_{|\underline{m}-\underline{m}_0|\leq V} |S(b_1;c_1\underline{m})| \cdot |S(b_2;c_1\underline{m})| \cdot |S(q_3;c_1\underline{m})| \\ &\ll q^{n/2+\varepsilon} b_1 b_2^2 D(b_1 b_2) \sum_{|\underline{m}-\underline{m}_0(\underline{m}_0)|\leq V} (\Phi(c_1\underline{m}), b_1)^{1/2} \times \\ &\qquad \sideset{}{^*}\sum_{\underline{a}}^{q_3} \#\mathrm{Null}_{q_3}(a_1M_1 + a_2M_2)^{1/2} \Delta_{T,q_3}(c_3\underline{m} + \underline{\mathfrak{b}})\end{aligned}$$

$$:= q^{n/2+\varepsilon} b_1 b_2^2 D(b_1 b_2) \sideset{}{^*}\sum_{\underline{a}}^{q_3} \#\mathrm{Null}_{q_3}(M(\underline{a}))^{1/2} B(b_1, q_3, V; \underline{m}_0) \tag{9.2}$$

where $M(\underline{a})$ be as in (8.18) and

$$B(b_1, q_3, V; \underline{m}_0) := \sum_{|\underline{m}-\underline{m}_0|\leq V} (\Phi(\underline{m}), b_1)^{1/2} \cdot \Delta_{T,q_3}(\underline{m} + \underline{\mathfrak{b}}'). \tag{9.3}$$

where $\underline{\mathfrak{b}}' \equiv c_3^{-1}\underline{\mathfrak{b}} \mod q_3$. We used $(\Phi(\underline{m}), b_1)$ instead of $(\Phi(c_1\underline{m}), b_1)$ in the definition of $B(b_1, q_3, V; \underline{m}_0)$ because Φ is homogeneous and $(b_1, c_1) = 1$. Likewise, by inspecting the definition of Δ, we can use $\Delta_{T,q_3}(\underline{m} + \underline{\mathfrak{b}}')$ in the definition of $B(b_1, q_3, V; \underline{m}_0)$ instead of $\Delta_{T,q_3}(c_3\underline{m} + \underline{\mathfrak{b}})$ since we can "divide through" by c_3, as $(c_3, q_3) = 1$ (in particular $(c_3, \lambda) = 1$ for any divisor, λ, of q_3).

The first and most difficult task for this section is to bound $B(b_1, q_3, V; \underline{m}_0)$. This will be quite a delicate task since we need to save over the $\underline{m}$ sum in two different ways simultaneously. The situation that we find ourselves in is somewhat similar to that of [21, Section 7], and will in principle follow the same kind of argument. However we will have to work significantly harder to attain a suitable bound here because in [21], the authors did not have the $\Delta_{T,q_3}(\underline{m} + \underline{\mathfrak{b}}')$ term in their equivalent of (9.3). We could just “remove" this term by bounding it from above by 1 and then use the argument in [21, Section 7] directly, but this will cause our final bound for (9.1) to be very bad when $\#Null_{q_3}(a_1M_1 + a_2M_2)$ is large.

The following Lemma will provide our main estimate for (9.3):

Lemma 9.1. *Let $b_1, q_3, V \in \mathbb{N}$ and $\underline{m}_0 \in \mathbb{Z}^n$. Furthermore, let c and q_3 be defined as follows:*

$$\hat{q}_3 := \prod_{\substack{p^e || q_3 \\ 2\nmid e}} p, \quad q_3 = c^2\hat{q}_3 \tag{9.4}$$

Then

$$B(b_1, q_3, V; \underline{m}_0) \ll b_1^{\varepsilon}\Big(b_1^{1/2}c^{n/2} + V^{n-1}b_1^{1/2}c^{1/2} + V^n\Big)\#\mathrm{Null}_c(M(\underline{a}))^{-1}.$$

Proof. We begin by noting that

$$(T^t\underline{x})_i \equiv 0 \mod (q_3, \lambda_i) \implies (T^t\underline{x})_i \equiv 0 \mod (c, \lambda_i),$$

and so by the definition of Δ_{T,q_3} (8.24) we clearly have that

$$\Delta_{T,q_3}(\underline{x}) = 1 \implies \Delta_{T,c}(\underline{x}) = 1$$

for any $\underline{x} \in \mathbb{Z}^n$. Therefore – since we are looking for an upper bound of $B(b_1, q_3, V; \underline{m}_0)$ – we may replace $\Delta_{T,q_3}(\underline{m} + \underline{\mathfrak{b}}')$ in (9.3) with $\Delta_{T,c}(\underline{m} + \underline{\mathfrak{b}}')$. Furthermore, since all elements of our sum are non-negative, we may extend the sum in (9.3) if we wish. In particular, the following bound must be true:

$$B(b_1, q_3, V; \underline{m}_0) \leq \sum_{|\underline{m}-\underline{m}_0| \leq \hat{V}} (\Phi(\underline{m}), b_1)^{1/2} \cdot \Delta_{T,c}(\underline{m} + \underline{\mathfrak{b}}'), \tag{9.5}$$

where

$$\hat{V} := \max\{V, c\}. \tag{9.6}$$

We have extended the sum up to $\hat{V}$ so that we can consider complete sums modulo c, as this will make it easier to acquire saving from $\Delta_{T,c}$ later. To this end, let $\underline{m} := \underline{m}_0 + \underline{m}_1 + c\underline{m}_2$, where $\underline{m}_1 \in (\mathbb{Z}/c\mathbb{Z})^n$ and $|\underline{m}_2| \leq \hat{V}/c$. Applying this decomposition on the right-hand side of (9.5) gives

$$\begin{aligned} B(b_1, q_3, V; \underline{m}_0) &\leq \sum_{\underline{m}_1 \bmod c} \sum_{|\underline{m}_2| \leq \hat{V}/c} (\Phi(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2), b_1)^{1/2} \times \\ &\qquad \Delta_{T,c}(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2 + \underline{\mathfrak{b}}') \\ &= \sum_{\underline{m}_1 \bmod c} \Delta_{T,c}(\underline{m}_0 + \underline{m}_1 + \underline{\mathfrak{b}}') \sum_{\underline{m}_2 \in U(\underline{m}_1)} (\Phi(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2), b_1)^{1/2}. \end{aligned} \tag{9.7}$$

The upshot of reordering our sum in this way is that we have managed to separate $\Delta_{T,c}(\underline{m}_0 + \underline{m}_1 + \underline{\mathfrak{b}}')$ and $(\Phi(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2), b_1)^{1/2}$. In particular, we can treat $\underline{m}_1$ as fixed for now, and since $\underline{m}_0$ and c are also fixed, we may focus on acquiring saving

in the $\underline{m}_2$ sum via $(\Phi_{c,\underline{m}_1}(\underline{m}_2), b_1)^{1/2}$, where

$$\Phi_{c,\underline{m}_0,\underline{m}_1}(\underline{m}_2) := \Phi(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2).$$

We observe that $(\Phi_{c,\underline{m}_0,\underline{m}_1}(\underline{m}_2), b_1)$ must be equal to some divisor of b_1, so we will decompose the $\underline{m}_2$ sum as follows:

$$\sum_{|\underline{m}_2| \leq \hat{V}/c} (\Phi(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2), b_1)^{1/2} = \sum_{d|b_1} d^{1/2} \#\{|\underline{x}| \leq \hat{V}/c : \Phi(\underline{m}_0 + \underline{m}_1 + c\underline{x}) \equiv 0 \mod d\}. \tag{9.8}$$

We now aim to use [3, Lemma 4] to bound the right hand side. Since Φ is homogeneous with $\mathrm{Cont}(\Phi) = 1$ by Proposition 8.5, and since c and d are co-prime, we have that

$$(\mathrm{Cont}(\Phi_{c,\underline{m}_0,\underline{m}_1}), d) \leq (\mathrm{Cont}(\Phi^{(0)}_{c,\underline{m}_0,\underline{m}_1}), d) = (c^{\deg(\Phi)}\mathrm{Cont}(\Phi), d) = (c^{\deg(\Phi)}, d) = 1.$$

Hence for every prime p dividing d, $\Phi(\underline{m}_0 + \underline{m}_1 + c\underline{x})$ is a non-trivial polynomial and therefore the corresponding variety is of dimension $n-1$. Therefore, we may now use [3, Lemma 4] to conclude that

$$\#\{|\underline{x}| \leq \hat{V}/c : \Phi(\underline{m}_0 + \underline{m}_1 + c\underline{x}) \equiv 0 \mod d\} \ll 1 + \left(\frac{V}{c}\right)^{n-1} + \left(\frac{V}{c}\right)^{n} d^{-1}.$$

Substituting this back into (9.8) gives the following:

$$\begin{aligned}\sum_{|\underline{m}_2| \leq \hat{V}/c} (\Phi(\underline{m}_0 + \underline{m}_1 + c\underline{m}_2), b_1)^{1/2} &\ll \sum_{d|b_1} d^{1/2} + \left(\frac{V}{c}\right)^{n-1} d^{1/2} + \left(\frac{V}{c}\right)^{n} d^{-1/2} \\ &\ll b_1^{\varepsilon}\left(b_1^{1/2} + \left(\frac{V}{c}\right)^{n-1} b_1^{1/2} + \left(\frac{V}{c}\right)^{n}\right). \end{aligned} \tag{9.9}$$

This in turn will enable us to find a suitable bound for $B(b_1, q_2, V; \underline{m}_0)$. By (9.7) and (9.9), we have

$$B(b_1, q_3, V; \underline{m}_0) \ll b_1^{\varepsilon}\left(b_1^{1/2} + \left(\frac{V}{c}\right)^{n-1} b_1^{1/2} + \left(\frac{V}{c}\right)^{n}\right) \sum_{\underline{x} \bmod c} \Delta_{T,c}(\underline{x} + \underline{m}_0 + \underline{\mathfrak{b}}'). \tag{9.10}$$

In order to find the bound we desire for $B(b_1, q_3, V; \underline{m}_0)$, we will need to turn our attention to the sum of type

$$\sum_{\underline{x} \bmod c} \Delta_{T,c}(\underline{x} + \underline{l}),$$

for some fixed $\underline{l} \in \mathbb{Z}^n$. Our bound here will be independent of the choice of the vector $\underline{l}$. This sum is much easier to handle since we have a complete sum at hand. It is easy to check from the definition of $\Delta_{T,c}(\underline{x} + \underline{l})$ (and the fact that $\det(T^t) = 1$) that

$$\begin{aligned}
\sum_{\underline{x} \bmod c} \Delta_{T,c}(\underline{x} + \underline{l}) &= \#\{\underline{x} \bmod c \,:\, (T^t\underline{x})_i \equiv -(\underline{l})_i \bmod \lambda_{c,i},\ i \in \{1, \cdots, n\}\} \\
&\leq \#\{\underline{x} \bmod c \,:\, (T^t\underline{x})_i \equiv 0 \bmod \lambda_{c,i},\ i \in \{1, \cdots, n\}\} \\
&= \#\{\underline{x} \bmod c \,:\, x_i \equiv 0 \bmod \lambda_{c,i},\ i \in \{1, \cdots, n\}\} \\
&= \frac{c^n}{\prod_i \lambda_{c,i}} \\
&= c^n \#\mathrm{Null}_c(M(\underline{a}))^{-1}.
\end{aligned}$$

Therefore, by (9.10), we have

$$B(b_1, q_3, V; \underline{m}_0) \leq b_1^{\varepsilon}\Big(b_1^{1/2}c^n + V^{n-1}b_1^{1/2}c + V^n\Big)\#\mathrm{Null}_c(M(\underline{a}))^{-1}, \tag{9.11}$$

as required. □

We can now ready to obtain a final bound for $\sum_{|\underline{m}-\underline{m}_0|\leq V} |S(q; \underline{m})|$. Before substituting (9.11) back into (9.2), we will perform some simplifications. Firstly, we note that by (9.4) and Lemma 4.5, we have

$$\#\mathrm{Null}_{q_3}(M(\underline{a}))^{1/2} \leq \#\mathrm{Null}_c(M(\underline{a}))\#\mathrm{Null}_{\hat{q}_3}(M(\underline{a}))^{1/2}. \tag{9.12}$$

Furthermore, by Proposition 4.6, we have

$$\begin{aligned}
\sum_{\underline{a}}^{q_3 *} \#\mathrm{Null}_{\hat{q}_3}(M(\underline{a}))^{1/2} &\leq \sum_{\underline{a}}^{q_3 *} \#\mathrm{Null}_{\hat{q}_3}(M(\underline{a})) \\
&\ll q_3^{2+\varepsilon} \prod_{p_i | \hat{q}_3} p_i^{m_{p_i}(F,G)+1}
\end{aligned}$$

$$= q_3^{2+\varepsilon} D(\hat{q}_3). \tag{9.13}$$

Finally, by combining (9.11)-(9.13) with (9.2), we arrive at the following bound:

Lemma 9.2. *For every $q \in \mathbb{N}$, if $n > 1$, then*

$$\sum_{|\underline{m}-\underline{m}_0|\leq V} |S(q;\underline{m})| \ll q^{1+n/2+\varepsilon} b_2 q_3 D(b_1 b_2 \hat{q}_3)\Big(b_1^{1/2} c^n + V^{n-1} b_1^{1/2} c + V^n\Big),$$

where $q_3 = c^2\hat{q}_3$ as defined in the statement of Lemma 9.1.

Recall that our ultimate goal was to find a suitable bound for $|T(q,\underline{z})|$. Upon noting that the above treatment of $\sum_{|\underline{m}-\underline{m}_0|\leq V} |S(q;\underline{m})|$ works for any value of $\underline{y} \in P\,\mathrm{Supp}(\omega)$ we may now substitute the bound in Lemma 9.2 into (9.1) to get the following bound for $T(q,\underline{z})$:

$$|T(q,\underline{z})| \ll 1 + P^n q^{1-n/2+\varepsilon} b_1^{1/2} b_2 q_3 D(b_1 b_2 \hat{q}_3)\Big(V^n b_1^{-1/2} + V^{n-1} c + c^n\Big)$$

If q is sufficiently small ($q < P^2$ say), then the right hand term dominates over 1 for every $n \geq 1$. Therefore, we finally reach the following bound for $|T(q,\underline{z})|$:

$$|T(q,\underline{z})| \ll P^n q^{1-n/2+\varepsilon} b_1^{1/2} b_2 q_3 D(b_1 b_2 \hat{q}_3)\Big(V^n b_1^{-1/2} + V^{n-1} c + c^n\Big),$$

where $q_3 = c^2\hat{q}_3$ as defined in Lemma 9.1. Note that if we use a weaker bound $c \leq b_3^{1/3} q_4^{1/2}$ and use the quality $q_3 = b_3 q_4$, the above bound becomes:

Proposition 9.3. *For every $q < P^2$, $\underline{z}$, and every $\varepsilon > 0$, if $n > 1$, we have*

$$|T(q,\underline{z})| \ll P^n q^{1-n/2+\varepsilon} b_1^{1/2} b_2 q_3 D(b_1 b_2 \hat{q}_3)\Big(V^n b_1^{-1/2} + V^{n-1} b_3^{1/3} q_4^{1/2} + b_3^{n/3} q_4^{n/2}\Big),$$

where n is the number of variables of F, G, b_3 is the 4th power-free cube part of q, and q_4 is the 4th power-full part of q.

The bound for the $n = 1$ case is much simpler to derive than in the $n > 1$ case. By

Lemma 8.4 and Propositions 8.6, 8.10, and 8.11, we have

$$\begin{aligned}\sum_{|m-m_0|\le V}|S(q;m)| &\ll q^{1/2+\varepsilon}b_1^{3/2}b_2^2D(b_1b_2)\times \\ &\qquad \sum_{\underline{a}}^{q_3{}^*}(q_3,M(\underline{a}))^{1/2}\sum_{|m-m_0|\le V}\Delta'_{q_3}(c_3m+\mathfrak{b}) \\ &\le q^{1/2+\varepsilon}b_1^{3/2}b_2^2D(b_1b_2)\times \\ &\qquad \sum_{\underline{a}}^{q_3{}^*}(q_3,M(\underline{a}))^{1/2}\sum_{|m-m_0|\le \max\{V,q_3\}}\Delta'_{q_3}(m+\mathfrak{b}') \\ &= q^{1/2+\varepsilon}b_1^{3/2}b_2^2D(b_1b_2)\sum_{\underline{a}}^{q_3{}^*}(q_3,M(\underline{a}))^{1/2}\Big(1+\frac{V}{(q,M(\underline{a}))}\Big) \\ &\le q^{1/2+\varepsilon}b_1^{3/2}b_2^2D(b_1b_2)\sum_{\underline{a}}^{q_3{}^*}\Big((q_3,M(\underline{a}))^{1/2}+V\Big). \qquad (9.14)\end{aligned}$$

We trivially have $\sum_{\underline{a}} V \le q_3^2V$. As for the other part of the sum, upon recalling that $q_3 = c^2\hat{q}_3$, we have

$$\begin{aligned}\sum_{\underline{a}}^{q_3{}^*}(q_3,M(\underline{a}))^{1/2} &\le c\sum_{\underline{a}}^{q_3{}^*}(\hat{q}_3,M(\underline{a}))^{1/2} \\ &\le c^5\sum_{\underline{a}}^{\hat{q}_3{}^*}(\hat{q}_3,M(\underline{a}))^{1/2} \\ &\ll q_3^2cD(\hat{q}_3),\end{aligned}$$

by the same argument as the proof of Proposition 8.11. Combining this with (9.14) gives the following result.

Lemma 9.4. *Let $q \in \mathbb{N}$, $m \in \mathbb{Z}$, $q_3 := c^2\hat{q}_3$ be defined as in Lemma 9.1. Then for every $\varepsilon > 0$,*

$$\sum_{|m-m_0|\le V}|S(q;m)| \ll q^{2+\varepsilon}(b_2q_3)^{1/2}D(b_1b_2\hat{q}_3)(V+c).$$

Finally, upon recalling that $q_3 = b_3q_4$, $c \le b_3^{1/3}q_3^{1/2}$, we may combine this lemma with (9.1) to get our final bound for $|T(q,\underline{z})|$ in the $n = 1$ case:

Proposition 9.5. *For every $q < P^2$, $\underline{z}$, and every $\varepsilon > 0$, if $n = 1$, we have*

$$|T(q,\underline{z})| \ll Pq^{1+\varepsilon}(b_2q_3)^{1/2}D(b_1b_2\hat{q}_3)(V+b_3^{1/3}q_4^{1/2}),$$

where n is the number of variables of F, G, b_3 is the 4th power-free cube part of q, and q_4 is the 4th power-full part of q.

Chapter 10

Finalisation of the Poisson bound

In this section, we will adapt the arguments used in [3, Section 7] and [21, Section 8] to our context in order to finalise our main bounds coming from Poisson summation. For a fixed value of t, Lemmas 7.1 and 7.5 allow us to consider bounding the sum

$$\sup_{t\ll|\underline{z}|\ll t}\sum_{\underline{h}\ll H}|T_{\underline{h}}(q,\underline{z})|,$$

where

$$T_{\underline{h}}(q,\underline{z}):=\sideset{}{^*}\sum_{\underline{a}\bmod q}\sum_{\underline{x}\in\mathbb{Z}^n}\omega_{\underline{h}}(\underline{x}/P)e((a_1/q+z_1)F_{\underline{h}}(\underline{x})+(a_2/q+z_2)G_{\underline{h}}(\underline{x}))$$

is the quadratic exponential sum as defined in (7.13). We may therefore apply our bounds for quadratic exponential sums in Propositions 9.3 and 9.5 to estimate these.

Now that $\underline{h}$ is allowed to vary, we will define

$$m_p(\underline{h}):=m_p(F_{\underline{h}}^{(0)},G_{\underline{h}}^{(0)}),\tag{10.1}$$

where $F_{\underline{h}}^{(0)}$ and $G_{\underline{h}}^{(0)}$ denote the leading quadratic parts of $F_{\underline{h}}$ and $G_{\underline{h}}$ respectively. We recall that $q=b_1b_2q_3$, where q_3 is the cube-full part of q, and b_1,b_2 are the square-free and cube-free square parts of q. Since we are fixing q for now, b_1, b_2, and q_3 are also fixed. Recall that we may write $b_i=b_{i,0}b_{i,1}\cdots b_{i,n}$, $q_3=q_{3,0}q_{3,1}\cdots q_{3,n}$

where $b_{i,j}$, and $q_{3,j}$ now depend on $\underline{h}$ and are defined to be

$$b_{i,j}(\underline{h}) := \prod_{\substack{p^i||b_i \\ m_p(\underline{h})=j-1}} p^i, \quad q_{3,j}(\underline{h}) := \prod_{\substack{p^e||q_3 \\ m_p(\underline{h})=j-1}} p^e.$$

We see that for any q fixed, there are at most $O(q^\varepsilon) = O(P^\varepsilon)$ possible choices for

$$\underline{c} = (b_{1,0}, \cdots, b_{1,n}, b_{2,0}, \cdots, b_{2,n}, q_{3,0}, \cdots, q_{3,n})$$

since there are only at most $O(q^\varepsilon)$ partitions of q into multiplicative factors. Therefore using the triangle inequality, we have that

$$\sum_{\underline{h} \ll H} |T_{\underline{h}}(q,z)| \leq P^\varepsilon \max_{\underline{c}} \Big\{ \sum_{\substack{\underline{h} \\ \underline{c}(\underline{h})=\underline{c}}} |T_{\underline{h}}(q,z)| \Big\} = P^\varepsilon \sum_{\substack{\underline{h} \\ \underline{c}(\underline{h})=\underline{c}'}} |T_{\underline{h}}(q,z)|$$

for some particular $\underline{c}'$, and $\underline{c}(\underline{h}) := (b_{1,0}(\underline{h}), \cdots, q_{3,n}(\underline{h}))$. We can then decompose this sum further by grouping $\underline{h}$'s with $m_\infty(\underline{h}) = s$:

$$\implies \sum_{\underline{h} \ll H} |T_{\underline{h}}(q,z)| \leq P^\varepsilon \sum_{s=-1}^{n-1} \sum_{\underline{h} \in \mathcal{H}_s} |T_{\underline{h}}(q,z)|, \tag{10.2}$$

where

$$\mathcal{H}_s := \{\underline{h} \in \mathbb{Z}^n : \underline{h} \ll H,\ \underline{c}'(\underline{h}) = \underline{c}',\ m_\infty(\underline{h}) = s\}. \tag{10.3}$$

Here, given v either a prime, or ∞ we define

$$m_v(\underline{h}) = \max\Big\{ s_v(F_{\underline{h}}^{(0)}), s_v(G_{\underline{h}}^{(0)}), s_v(F_{\underline{h}}^{(0)}, G_{\underline{h}}^{(0)}) \Big\}. \tag{10.4}$$

We now aim to estimate the size of $\mathcal{H}_s$. We start by noting that we must have that $\mathcal{H}_s = \emptyset$ unless $b_{1,i} = b_{2,i} = q_{3,i} = 1$ for $i \leq s$. This is because $m_p(\underline{h}) \geq m_\infty(\underline{h})$ for every p. To get a bound on $\#\mathcal{H}_s$ we will start by constructing a set which contains $\mathcal{H}_s$ that is easier to work with. Let

$$V_{v,i} := \{\underline{h} \in \mathbb{A}^n_{\overline{\mathbb{F}}_v} \mid m_v(\underline{h}) \geq i - 1\}$$

Then, upon defining $[\underline{h}]_p$ to be the reduction modulo p of a point $\underline{h} \in \mathbb{Z}^n$, we have

$$\mathcal{H}_s \subset \{\underline{h} \in V'_{\infty,s+1} \cap \Big[-H, H\Big]^n | [\underline{h}]_p \in V_{p,i} \text{ for all } p | b_{1,i} b_{2,i} q_{3,i}\}. \tag{10.5}$$

In order to bound this larger set, we will need the following lemma, which is analogous to [21, Lemma 8.2].

Lemma 10.1. *Let*

$$m_\infty := \max\{s_\infty(F), s_\infty(G), s_\infty(F,G)\},$$

and let $V_{v,i}$ be a closed subvariety in $\mathbb{A}^n_{\overline{\mathbb{F}}_v}$ of degree $O(1)$, and there is an absolute constant C s.t.

$$\dim(V_{v,i}) \leq \min\{n, n + m_\infty + 1 - i\}$$

as long as $v = p > C$ or $v = \infty$.

Proof. Since

$$m_v(\underline{h}) = \max\left\{s_v(F^{(0)}_{\underline{h}}), s_v(G^{(0)}_{\underline{h}}), s_v(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}})\right\},$$

we can write $V_{v,i}$ as the union of three sets $V_{v,i,j}$ for $j = 1, 2, 3$ defined by

$$s_v(F^{(0)}_{\underline{h}}) \geq i - 1, \quad s_v(G^{(0)}_{\underline{h}}) \geq i - 1, \quad s_v(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}}) \geq i - 1$$

respectively. By following the same argument as in [3][Lemma 1], we can conclude that $\max\{\dim(V_{v,i,1}), \dim(V_{v,i,2})\} \leq \min\{n, n + m_\infty + 1 - i\}$. Moreover, since

$$s_v(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}}) = \dim(\{\underline{x} \in \mathbb{P}^{n-1}_{\overline{\mathbb{F}}_v} \mid \underline{h} \cdot \nabla F^{(0)}(\underline{x}) = \underline{h} \cdot \nabla G^{(0)}(\underline{x}) = 0, \operatorname{Rank}\begin{pmatrix} \underline{h} \cdot \nabla^2 F^{(0)}(\underline{x}) \\ \underline{h} \cdot \nabla^2 G^{(0)}(\underline{x}) \end{pmatrix} < 2\}),$$

then we can use [22, Lemma 3(ii)] to conclude that $\dim(V_{v,i,3}) \leq \min\{n, n + m_\infty + 1 - i\}$, provided that $v = p \gg 1$. Therefore we only need to check $V_{\infty,i,3}$. We will use a slight modification to the argument used in [3, Lemma 1] in order to show that $\dim(V_{\infty,i,3}) \leq \min\{n, n + m_\infty + 1 - i\}$: Let

$$U(F,G) = U := \{(\underline{x}, \underline{y}) \in \mathbb{A}^{2n}_{\mathbb{Q}} \mid \underline{y} \cdot \nabla F(\underline{x}) = \underline{y} \cdot \nabla G(\underline{x}) = 0, \operatorname{Rank}\begin{pmatrix} \underline{y} \cdot \nabla^2 F(\underline{x}) \\ \underline{y} \cdot \nabla^2 G(\underline{x}) \end{pmatrix} < 2\}$$

for F, G homogeneous forms of degree 3, and let $D := \{(\underline{x}, \underline{y}) \in \mathbb{A}^{2n}_{\mathbb{Q}} \mid \underline{x} = \underline{y}\}$. Then

by the Affine Dimension Theorem, we have that

$$\dim(U) \leq \dim(U \cap D) - \dim(D) + 2n = \dim(U \cap D) + n. \tag{10.6}$$

Next, we note that

$$U \cap D = \{\underline{x} \in \mathbb{A}^n_{\mathbb{Q}} \mid \underline{x} \cdot \nabla F(\underline{x}) = \underline{x} \cdot \nabla G(\underline{x}) = 0,\ \text{Rank} \begin{pmatrix} \nabla(\underline{x} \cdot \nabla F(\underline{x})) \\ \nabla(\underline{x} \cdot \nabla G(\underline{x})) \end{pmatrix} < 2\}$$

$$= \{\underline{x} \in \mathbb{A}^n_{\mathbb{Q}} \mid F(\underline{x}) = G(\underline{x}) = 0,\ \text{Rank} \begin{pmatrix} \nabla F(\underline{x}) \\ \nabla G(\underline{x}) \end{pmatrix} < 2\}$$

by Euler's identity. Hence, by (10.6), we have

$$\dim(U \cap D) = s_\infty(F, G) + 1$$

and so

$$\dim(U) \leq n + s_\infty(F, G) + 1 \leq n + m_\infty + 1. \tag{10.7}$$

Finally we let $F = F^{(0)}$, $G = G^{(0)}$. If

$$\dim(V_{\infty,i,3}) > n + m_\infty + 1 - i,$$

then, by definition we have that

$$\dim(\{(\underline{x}, \underline{h}) \in \mathbb{A}^{2n}_{\mathbb{Q}} \mid s_\infty(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}}) \geq i - 1,\ \underline{x} \in s_\infty(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}})\}) > (n + m_\infty + 1 - i) + i$$

$$= n + m_\infty + 1.$$

It is easy to check that

$$\{(\underline{x}, \underline{h}) \in \mathbb{A}^{2n}_{\mathbb{Q}} \mid s_\infty(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}}) \geq i - 1,\ \underline{x} \in s_\infty(F^{(0)}_{\underline{h}}, G^{(0)}_{\underline{h}})\} \subset U((F^{(0)}, G^{(0)})),$$

and so

$$\dim(U(F^{(0)}, G^{(0)})) > n + m_\infty + 1.$$

This contradicts (10.7). Hence $\dim(V_{\infty,i,3}) \leq n + m_\infty + 1 - i$ as required. □

We can now use (10.5) and the argument found in [3, Section 7] to get the following upper bound for $\#\mathcal{H}$:

$$\#\mathcal{H}_s \ll q^{\varepsilon} \max_{s+1\leq\eta\leq n} \frac{H^{n-\eta}}{\prod_{i=\eta+1}^{n}(b_{1,i}b_{2,i}^{1/2}\tilde{q}_{3,i})^{i-\eta}}, \tag{10.8}$$

where

$$\tilde{q}_{3,i} := \prod_{\substack{p|q_3 \\ m_p(\underline{h})=i-1}} p^i.$$

For convenience set

$$\mathcal{U}_s := \sum_{\underline{h}\in\mathcal{H}_s} T_{\underline{h}}(q,\underline{z}) \tag{10.9}$$

(recall that $\sum_{\underline{h}\ll H} T_{\underline{h}}(q,\underline{z}) \ll P^{\varepsilon}\sum_{s=-1}^{n-1}\mathcal{U}_s$ by (10.2)). We will use (10.8) to bound $\mathcal{U}_s$ later, but for now, we need to find a bound on $|T_{\underline{h}}(q,\underline{z})|$. To do this we will need to apply the hyperplane intersections lemma, namely Lemma 4.2 and then apply the bounds found in Propositions 9.3 and 9.5.

Let η be chosen so as to maximize the expression in (10.8). Let Π be the set of primes $p|q$ so that $r=\omega(q)$, and $\{F_1,F_2\}=\{F_{\underline{h}}^{(0)},G_{\underline{h}}^{(0)}\}$. We may now invoke Lemma 4.2 to find a lattice Λ_η of rank $n-\eta$ and a basis $\underline{e}_1,\cdots,\underline{e}_{n-\eta}$ for Λ_η s.t. for every $\underline{t}\in\mathbb{Z}^n$, the polynomials

$$\tilde{F}_{\underline{h},\underline{t}}(\underline{y}) := F_{\underline{h}}^{(0)}(\underline{t}+\sum_{i=1}^{n-\eta} y_i\underline{e}_i), \quad \tilde{G}_{\underline{h},\underline{t}}(\underline{y}) := G_{\underline{h}}^{(0)}(\underline{t}+\sum_{i=1}^{n-\eta} y_i\underline{e}_i)$$

satisfy

$$m_v(\tilde{F}_{\underline{h},\underline{t}},\tilde{G}_{\underline{h},\underline{t}}) = \max\{-1, m_v(F_{\underline{h}}^{(0)},G_{\underline{h}}^{(0)})-\eta\} \tag{10.10}$$

for every $v\in\{\infty\}\cup\Pi_{cr}$. We also note that $\deg(\tilde{F}_{\underline{h},\underline{t}})=\deg(\tilde{G}_{\underline{h},\underline{t}})=2$ (this is necessary in order to be able to use the bounds from the previous chapter). In order to apply the bounds found in the previous chapter, we must first fix our choice of basis $\{\underline{e}_1,\cdots,\underline{e}_n\}$, and so we will use the same process as earlier when we fixed $(b_{1,0},\cdots,q_{4,n})$: We recall that the L used in (4.5) is of size $L=O(r+1)=O(\log(q))$. Therefore there are at most $O(\log(q)^n)$ choices of basis satisfying (4.5), and so by

(10.9), and the triangle inequality, there is one such choice for which

$$\mathcal{U}_s \ll \log(q)^n \sum_{\underline{h}\in\mathcal{H}_s}{}' |T_{\underline{h}}(q,\underline{z})| \ll P^{\varepsilon} \sum_{\underline{h}\in\mathcal{H}_s}{}' |T_{\underline{h}}(q,\underline{z})|, \tag{10.11}$$

where $\sum'$ denotes that the sum is taken over the vectors $\underline{h}$ in the original sum for which (10.10) holds for our chosen basis $\{\underline{e}_1,\cdots,\underline{e}_n\}$. For such $\underline{h}$, we can now separate the $\underline{x}$ sum defining $T_{\underline{h}}(q,\underline{z})$ into cosets $\underline{t}+\Lambda_\eta$ of Λ_η, where $\underline{t}$ runs over some subset $T_\eta \subset \mathbb{Z}^n$. All that is left to do is use Proposition 9.3 (or Proposition 9.5 for $\eta = n-1$) on each coset, and determine the size of T_η, as this bounds the number of cosets that we have. We claim that if Λ_η is chosen according to Lemma 4.2, then $\#T_\eta = O(P^\eta)$. Indeed, consider $\underline{x}$ in terms of our basis $\underline{e}_1,\cdots,\underline{e}_n$, i.e. writing

$$\underline{x} = \sum_{i=1}^{n} u_i \underline{e}_i.$$

Now, if π_i denotes the projection onto the orthogonal subspace spanned by the vectors $\underline{e}_j$, $i \neq j$, we have

$$||\underline{x}|| \geq ||\pi_i \underline{x}|| = |u_i| \cdot ||\pi \underline{e}_i|| = |u_i| \frac{|\det(\Lambda)|}{|\det(\Lambda_i)|},$$

where $\Lambda \subset \mathbb{Z}^n$ denotes the full-dimensional lattice spanned by $\underline{e}_1,\cdots,\underline{e}_n$ and Λ_i the lattice spanned by each $\underline{e}_j \neq \underline{e}_i$. Now by (4.5) and (4.6), we get that

$$|u_i| \ll \frac{||\underline{x}||}{L}. \tag{10.12}$$

Therefore we certainly have $|u_i| \ll P$ since we need $||\underline{x}|| \ll P$. Hence, since $\Lambda_\eta =< \underline{e}_1,\cdots,\underline{e}_{n-\eta} >$, we may conclude that $\underline{t}$ is of the form $\underline{t} = \sum_{i=n-\eta+1}^{n} \lambda_i \underline{e}_i$ s.t. $|\lambda_i| \ll P$. We now choose T_η to be the collection of such $\underline{t}$ leading us to conclude that $\#T_\eta = O(P^\eta)$.

In order to complete the hyperplane intersections step, we will now define new weight functions in $n-\eta$ variables. In particular, we set

$$\tilde{\omega}_{\underline{h},\underline{t}}(y_1,\cdots,y_{n-\eta}) := \omega_{\underline{h}}\Big(P^{-1}\underline{t} + L^{-1}\sum_{i=1}^{\eta} y_i \underline{e}_i\Big).$$

This gives us

$$|T_{\underline{h}}(q,\underline{z})| \leq \sum_{\underline{t}\in T_\eta} |T_{\underline{h},\underline{t}}(q,\underline{z})|, \text{ where } T_{\underline{h},\underline{t}}(q,\underline{z}) = T_{n-\eta}(q,\underline{z};\tilde{F}_{\underline{h},\underline{t}},\tilde{G}_{\underline{h},\underline{t}},\tilde{\omega}_{\underline{h},\underline{t}},P/L) \tag{10.13}$$

We now need to verify that $T_{\underline{h},\underline{t}}(q,\underline{z})$ and $\tilde{\omega}_{\underline{h},\underline{t}}$ satisfy the various properties that we assumed in order to acquire the results we have found in the previous sections. Firstly, we refer to the proof Proposition 2 of [3] to see that $\tilde{\omega}_{\underline{h},\underline{t}} \in \mathcal{W}_{n-\eta}$ for $\underline{t} \ll P$. We also see that

$$||\tilde{F}_{\underline{h},\underline{t}}||_{P/L} \ll L^2||F_{\underline{h}}||P \ll P^\varepsilon H||F||_p \ll P^\varepsilon H,$$

and similarly $||\tilde{G}_{\underline{h},\underline{t}}||_{P/L} \ll P^\varepsilon H$. Next, we note that $\eta \geq s+1$, and so we automatically have $m_\infty = -1$. This covers all conditions that we have needed in the previous sections on exponential sums.

Therefore, by (10.11), (10.13), and (10.8):

$$\begin{aligned}
\mathcal{U}_s &\ll P^\varepsilon {\sum_{\underline{h}\in\mathcal{H}_s}}' \sum_{\underline{t}\in T_\eta} |T_{\underline{h},\underline{t}}(q,\underline{z})| \\
&\ll P^\varepsilon \#\mathcal{H}_s \#T_\eta {\max_{\underline{h}\in\mathcal{H}_s}}' \max_{\underline{t}\in T_\eta} |T_{\underline{h},\underline{t}}(q,\underline{z})| \\
&\ll \max_{s+1\leq\eta\leq n} \frac{P^{\eta+\varepsilon}H^{n-\eta}}{\prod_{i=\eta+1}^n((b_{1,i}b_{2,i}^{1/2}\tilde{q}_{3,i}))^{i-\eta}} \cdot {\max_{\underline{h}\in\mathcal{H}_s}}' \max_{\underline{t}\in T_\eta} T_{\underline{h},\underline{t}}(q,\underline{z})
\end{aligned} \tag{10.14}$$

Recall that

$$\sum_{\underline{h}\ll H} T_{\underline{h}}(q,\underline{z}) \ll P^\varepsilon \sum_{s=-1}^{n-1} \mathcal{U}_s \ll P^\varepsilon \max_{-1\leq s\leq n-1} \mathcal{U}_s \tag{10.15}$$

by (10.2) and (10.9). We will therefore be able to attain our final bound for $\sum_{\underline{h}\ll H} T_{\underline{h}}(q,\underline{z})$ if we can find a bound for $T_{\underline{h},\underline{t}}(q,\underline{z})$.

We may use Propositions 9.3 and 9.5 to bound $T_{\underline{h},\underline{t}}(q,\underline{z})$ from above when $\eta < n-1$ and $\eta = n-1$ respectively. When $\eta = n$, we may proceed by a much simpler argument to bound $T_{\underline{h},\underline{t}}(q,\underline{z})$. We trivially have

$$|T_{\underline{h}}(q,\underline{z})| \leq {\sum_{\underline{a}}}^* \sum_{\underline{y}\in\mathbb{Z}^n} \omega_{\underline{h}}(\underline{y}/P) \ll q^2P^n,$$

and by Lemma 10.1 ($v = \infty$, $i = n$), we have that

$$\#\{\underline{h} \in \mathbb{A}_{\mathbb{Q}}^n \mid m_\infty(\underline{h}) = n-1\} = O(1).$$

Hence

$$\sum_{\substack{\underline{h} \ll H \\ m_\infty(\underline{h})=n-1}} |T_{\underline{h}(q,\underline{z})}| \ll q^2 P^n. \tag{10.16}$$

Returning to $\eta \leq n-1$: By (10.10), we may use the proof of Proposition 2 from [3] to conclude that for every $\underline{t} \in T_\eta$, we have

$$D_{\tilde{F}_{\underline{h},\underline{t}},\tilde{G}_{\underline{h},\underline{t}}}(b_{1,i}b_{2,i}\tilde{q}_{3,i}) \ll q^\varepsilon \prod_{i=\eta+1}^{n} (b_{1,i}b_{2,i}^{1/2}\tilde{q}_{3,i})^{i-\eta}$$

when $\eta < n-1$. When $\eta = n-1$, $(p, \mathrm{Cont}(\tilde{F}_{\underline{h},\underline{t}}), \mathrm{Cont}(\tilde{G}_{\underline{h},\underline{t}})) = p$ if and only if $p|\tilde{F}_{\underline{h},\underline{t}}, \tilde{G}_{\underline{h},\underline{t}}$ or $p \ll P^\varepsilon$. In particular, $p|b_{1,n}b_{2,n}^{1/2}\tilde{q}_{3,n}$ or $p \ll P^\epsilon \asymp q^\varepsilon$, and so we again have

$$D_{\tilde{F}_{\underline{h},\underline{t}},\tilde{G}_{\underline{h},\underline{t}}}(b_{1,n}b_{2,n}\tilde{q}_{3,n}) \ll q^\epsilon b_{1,n}b_{2,n}^{1/2}\tilde{q}_{3,n}.$$

Therefore, by (10.14) (10.15), and Propositions 9.3 and 9.5 and (10.16), we may conclude the following:

Proposition 10.2. *Let $q < P^2$, and let*

$$\mathcal{Y}_\eta := \frac{H^{n-\eta}}{q^{(n-\eta)/2}} b_1^{-1}\Big(V^{n-\eta} + V^{n-\eta-1}b_1^{1/2}b_3^{1/3}q_4^{1/2} + b_1^{1/2}b_3^{(n-\eta)/3}q_4^{(n-\eta)/2}\Big)$$

for $\eta \in \{0, \cdots, n-2\}$,

$$\mathcal{Y}_{n-1} := \frac{H}{q^{1/2}} b_1^{-1/2}(V + b_3^{1/3}q_4^{1/2}).$$

Then

$$\sum_{\underline{h} \in H_s} |T_{\underline{h}}(q,\underline{z})| \ll q^2 P^{n+\varepsilon}\Big(1 + \sum_{\eta=0}^{n-1} \mathcal{Y}_\eta\Big).$$

Chapter 11

A Simple Algorithm for Piecewise Linear Functions

In the previous sections of this thesis, we have built up the framework to bound our minor arcs in several different ways. In particular, we may use van der Corput differencing (with or without averaging) followed by either Poisson summation or Weyl differencing to get a total of four different bounds for the minor arcs, and we may use Weyl differencing twice to get a fifth bound. In theory all that is left to do therefore is to explicitly state these bounds, and then perform the *minor arcs optimisation step*. That is, to show that (provided that n is sufficiently large) at least one of our bounds, bounds the minor arcs by $O(P^{n-6-\delta})$ for every $q, \underline{z}$ covered by $S_{\mathfrak{m}}$. This process has typically been performed by hand up until now, however in our case this is impractical due to our Poisson summation bounds being much harder to work with than usual.

Fortunately, it turns out that the optimisation process can be turned into a piecewise linear programming problem and so – assuming one derives and builds an appropriate algorithm – we may instead rely on a computer to show that the minor arcs are bounded by $O(P^{n-6-\delta})$. There are many algorithms to find the maximum of a piecewise linear function on a given domain in the literature (the Simplex Algorithm is a well known example), but most of them assume that the function is "separable",

which is not something that we can assume in this context.

It is quite possible that there is a suitable algorithm already in the literature, but the author is not aware of one that is easy to adapt to the minor arcs bounds. For that reason, we will instead derive a crude algorithm to find the maximum (or minimum) of an arbitrary continuous, piecewise linear function defined on some convex polytope, which will also be easy to apply to our minor arcs bounds.

Our starting point will be a generalisation to the Fundamental Theorem of Linear Programming (FTLP). Let D be a convex polytope – that is, a set defined by the intersection of a collection of half-spaces – and let $V(D)$ be its vertices which are defined as follows: $\underline{x} \in D$ is a vertex of D if for every line segment $L : [0,1] \to \mathbb{R}^m$ such that $L([0,1]) \subset D$ and $\underline{x} \in L([0,1])$, we either have $\underline{x} = L(0)$ or $\underline{x} = L(1)$.

Lemma 11.1 (Fundamental Theorem of Linear Programming)**.** *Let D be a convex, compact polytope, and let $f : D \to \mathbb{R}$ be a linear function. Then*

$$\max_{\underline{x} \in D} f(\underline{x}) = \max_{\underline{x} \in V(D)} f(\underline{x}),$$

$$\min_{\underline{x} \in D} f(\underline{x}) = \min_{\underline{x} \in V(D)} f(\underline{x}).$$

FTLP essentially tells us that the maximum (or minimum) or a linear function on a convex polytope is at one of its corners. To generalise this idea to piecewise functions, we know that at every point in our domain D, our piecewise linear function F will correspond to a linear function. So in theory, if we could split D into finitely many pieces such that F is linear on each piece, then we could apply FTLP to each piece to find the maximum (or minimum) of F on D. The formal statement of this idea is the following:

Corollary 11.2 (Generalisation to FTLP)**.** *Let $D \subset \mathbb{R}^m$, F be a piecewise linear function, and let $D_1, \cdots, D_l \subset D$ be convex, compact polytopes with the following properties:*

1. $\bigcup_{i=1}^{l} D_i = D$.

2. F is linear on D_i for every $i \in \{1, \cdots, l\}$.

Then

$$\max_{\underline{x} \in D} F(\underline{x}) = \max_{i \in \{1,\cdots,l\}} \max_{\underline{x} \in V(D_i)} F(\underline{x}),$$
$$\min_{\underline{x} \in D} F(\underline{x}) = \min_{i \in \{1,\cdots,l\}} \min_{\underline{x} \in V(D_i)} F(\underline{x}).$$

Proof. We will only prove "max" since the argument is identical for "min". Since F is linear on each D_i and each D_i is a convex polytope, we have

$$\max_{\underline{x} \in D_i} F(\underline{x}) = \max_{\underline{x} \in V(D_i)} F(\underline{x})$$

for every $i \in \{1, \cdots, l\}$ by Lemma 11.1. Hence by property 1 from the corollary, we have:

$$\begin{aligned} \max_{\underline{x} \in D} F(\underline{x}) &= \max_{i \in \{1,\cdots,l\}} \max_{\underline{x} \in D_i} F(\underline{x}) \\ &= \max_{i \in \{1,\cdots,l\}} \max_{\underline{x} \in V(D_i)} F(\underline{x}), \end{aligned}$$

as required. □

From now on, we will no longer mention finding the minimum value of F except when stating theorems since our arguments are not sensitive to whether we are finding $\min(F)$ or $\max(F)$. In order to use Corollary 11.2 to create a suitable algorithm, we will need to do two things: Firstly, we need to find a collection of convex polytopes which satisfy the conditions in Corollary 11.2, and then we need to find a set of points which contains the vertices of these polytopes.

We will therefore work towards defining a collections of sets and then showing that they have the desired properties: Let $\text{Func}(F) := \{f_1, \cdots, f_k\}$ be the set of functional values that F can take on D. Let $\underline{p} \in D$ be a point, and $L_{\underline{p},\underline{x}}$ be the line segment between $\underline{p}$ and another point $\underline{x}$. Then, we may define

$$D_i(\underline{p}) := \{\underline{x} \in D \ : \ F(L_{\underline{p},\underline{x}}[0,1]) = f_i(L_{\underline{p},\underline{x}}[0,1])\}.$$

In other words, $D_i(\underline{p})$ is the set of points which are connected to $\underline{p}$ by a line, which have the property that F coincides with the linear function f_i on the entire line. In the next few lemmas, we will show that a certain finite collection of these sets will have all the properties that we desire.

For now, we note that for "most" $\underline{p}$, the set $D_i(\underline{p})$ will be empty for all but one i since the converse of this would imply that $F(\underline{p})$ coincides with at least two linear functions simultaneously at $\underline{p}$. This would mean that $\underline{p} \in \{\underline{x} \,:\, f_{i_1}(\underline{x}) = f_{i_2}(\underline{x})\}$ for some $i_1 \neq i_2 \in \{1, \cdots, k\}$, and so $\underline{p}$ must lie on an $(m-1)$-dimensional hyperplane.

The purpose of defining $D_i(\underline{p})$ in this way is we get that F is linear on $D_i(\underline{p})$ for free. We now aim to find conditions on F and D which will allow us to conclude that $D_i(\underline{p})$ is a convex, compact polytope, and that we are able to decompose D into a finite union of these sets. If we can do this, then Corollary 11.2 will allow us to find the maxima or minima of a piecewise linear function on domain D by testing finitely many points. To this end, we will start by proving a necessary auxiliary lemma:

Lemma 11.3. *Let F be a continuous, piecewise linear function and let D be convex. Then, for every $\underline{p} \in D$, $i \in \{1, \cdots, k\}$, we have that $D_i(\underline{p})$ is path connected, and*

$$\partial D_i(\underline{p}) \subset \partial D \bigcup_{j \neq i}^{k} \{\underline{x} \,:\, f_j(\underline{x}) = f_i(\underline{x})\}.$$

Proof. This lemma is an exercise in elementary analysis, so we will be brief: $D_i(\underline{p})$ being path connected is simply by definition.

The statement about $\partial D_i(\underline{p})$ follows almost immediately from the fact that F is continuous: The case where $D_i(\underline{p}) = \emptyset$ is trivial, so we will assume that this isn't the case. Let $\underline{b} \in \partial D_i(\underline{p})$. Then there must be some $\underline{x} \in D_i(\underline{p})$, $\underline{y} \notin D_i(\underline{p})$ and some $\lambda \in [0, 1)$ such that $L_{\underline{x},\underline{y}}(\lambda) = \underline{b}$ and $L_{\underline{x},\underline{y}}(\lambda') \notin D_i(\underline{p})$ for every $\lambda \leq \lambda' \leq 1$. For convenience, set $L = L_{\underline{x},\underline{y}}$.

If $L(\lambda') \notin D$ for every $\lambda' > \lambda$, then $L(\lambda) \in \partial D$. Alternatively, if there is some $\varepsilon > 0$ such that $L([\lambda, \lambda + \varepsilon]) \subset D$, then we must have $F(\underline{u}) \neq f_i(\underline{u})$ for every $\underline{u} \in L((\lambda, \lambda + \varepsilon])$ by the definition of $D_i(\underline{p})$.

Therefore (upon choosing ε to be smaller if necessary), we have that $F(\underline{u}) = f_j(\underline{u})$ for some $j \neq i$, for every $\underline{u} \in L((\lambda, \lambda + \varepsilon])$. However, $F(L(\lambda)) = f_i(L(\lambda))$. Hence by continuity of F, we must have $f_i(\underline{x}) = f_j(\underline{x})$ at $L(\lambda) \in \partial D_i(\underline{p})$.

Finally, we note that $\underline{b}$ was chosen to be an arbitrary point of $\partial D_i(\underline{p})$, and so we may conclude that

$$\partial D_i(\underline{p}) \subset \partial D \bigcup_{j \neq i}^{k} \{\underline{x} \,:\, f_j(\underline{x}) = f_i(\underline{x})\},$$

as required. □

Lemma 11.3 tells us about the boundary of $D_i(\underline{p})$ which in turn gives us information about the shape of $D_i(\underline{p})$. It should therefore not be too surprising that we can use this information to verify that $D_i(\underline{p})$ is a convex polytope. This will be the covered in the next corollary.

Corollary 11.4. *Let D be a convex, compact polytope. Then $D_i(\underline{p})$ is also a convex, compact polytope. In particular, there exist some $S_-, S_+ \subset \{1, \cdots, k\}$ such that*

$$D_i(\underline{p}) = D_i(S_-, S_+)$$

where

$$D_i(S_-, S_+) := \{\underline{x} \in D \,:\, f_i(\underline{x}) \leq f_{j_1}(\underline{x}),\ j_1 \in S_-,\ f_i(\underline{x}) \geq f_{j_2}(\underline{x}),\ j_2 \in S_+\}.$$

Proof. We will start by considering

$$\hat{D}_i(\underline{p}) := \{\{\underline{x} \in \mathbb{R}^m \,:\, F(L_{\underline{p},\underline{x}}[0,1]) = f_i(L_{\underline{p},\underline{x}}[0,1])\}\}.$$

By Lemma 11.3, $\hat{D}_i(\underline{p})$ is path connected and its boundary is a subset of hyperplanes defined by equations of the form $f_i(\underline{x}) = f_j(\underline{x})$, where $f_j \in \text{Func}(F)$. Hence, $\hat{D}_i(\underline{p})$ must lie completely on one side of these hyperplanes. We define S_- and S_+ accordingly (Note: not every hyperplane of this type necessarily defines a part of the boundary and so $S_- \cup S_+$ does not necessarily equal $\{1, \cdots, k\}$).

We can therefore express $\hat{D}_i(\underline{p})$ as an intersection of a finite number of halfspaces, which is equivalent to saying that $\hat{D}_i(\underline{p})$ is a convex polytope. Finally, we note that

D is also a convex polytope, $D_i(\underline{p}) = \hat{D}_i(\underline{p}) \cap D$, and finite intersections of convex polytopes are convex polytopes. $D_i(\underline{p})$ is also compact due to D being compact, and $D_i(\underline{p}) \subset D$ therefore being closed by definition (closed subsets of compact spaces are also compact). □

We are now ready to apply Corollary 11.2:

Corollary 11.5. *Let D be a bounded convex polytope and F a continuous, piecewise linear function such that* $\text{Func}(F) = \{f_1, \cdots, f_k\}$. *Then there exists a finite collection of points $\underline{p}_j \in D$ and some $i_j \in \{1, \cdots, k\}$ such that*

$$\max_{\underline{x} \in D} F(\underline{x}) = \max_j \max_{\underline{x} \in V(D_{i_j}(\underline{p}_j))} F(\underline{x}),$$

$$\min_{\underline{x} \in D} F(\underline{x}) = \min_j \min_{\underline{x} \in V(D_{i_j}(\underline{p}_j))} F(\underline{x}).$$

Proof. We have that $D_i(\underline{p})$ is a convex polytope for every $\underline{p} \in D$ by Corollary 11.4. Corollary 11.4 also shows that there can only be finitely many possible distinct sets that $D_i(\underline{p})$ can be since $D_i(\underline{p}) = D(S_-, S_+)$ for some S_-, S_+, and there are only finitely many possible choices for S_- and S_+. Furthermore every $\underline{x} \in D$ must lie in some $D(S_-, S_+)$. Combining these two facts allows us to conclude that there must be some collection of points $\{\underline{p}_1, \cdots, \underline{p}_r\}$ and some $i_j \in \{1, \cdots, k\}$, $j \in \{1, \cdots, r\}$, such that

$$D = \bigcup_{j=1}^{r} D_{i_j}(\underline{p}_j).$$

Finally, F is linear on $D_{i_j}(\underline{p}_j)$ by definition, and so we apply Corollary 11.2 to reach the desired result. □

We have successfully shown that there is a finite collection of points that we need to test to determine min/max of F defined on some polytope D, however we are still not quite finished. In order to build a program to find min/max of F, we still need to determine what the set $V(D_{i_j}(\underline{p}_j))$ is. For the purposes of building an algorithm, it is sufficient to find a "not too large" collection of points which contains $V(D_{i_j}(\underline{p}_j))$. To this end, we will need some new definitions.

Since $D \subset \mathbb{R}^m$ is assumed to be a bounded, convex polytope of dimension m, we have that its boundary is a finite union of $(m-1)$-dimensional hyperplanes. Define $B_1, \cdots, B_d$ to be these hyperplanes. Furthermore, let $B_{i,j}$ be the hyperplane defined by the equation $f_i(\underline{x}) = f_j(\underline{x})$ $(f_i, f_j \in \text{Func}(F))$, and let $\text{Planes}_{D,F}(i)$ be the collection of all of these hyperplanes. i.e.

$$\text{Planes}_{D,F}(i) := \{B_{i,1}, \cdots, B_{i,i-1}, B_{i,i+1}, \cdots, B_{i,k}, B_1, \cdots, B_d\}.$$

Note that this set will have at least m distinct hyperplanes due to D being compact. Finally, let

$$\text{Critical}_{D,F}(i) := \bigcup_{\substack{P_j \in \text{Planes}_{D,F}(i) \\ j \in \{1,\cdots,m\} \\ \#P_1 \cap \cdots \cap P_m = 1}} P_1 \cap \cdots \cap P_m.$$

In other words, $\text{Critical}_{D,F}(i)$ is the set of points attained by intersecting any m hyperplanes from $\text{Planes}_{D,F}(i)$. The third condition in the union is to exclude empty intersections and choosing the same plane twice. We can now state our lemma:

Lemma 11.6. *Let D be a convex polytope and F a continuous, piecewise linear function where* $\text{Func}(F) = \{f_1, \cdots, f_k\}$. *Then for any $i \in \{1, \cdots, k\}$ and any $\underline{p} \in D$, we have*

$$V(D_i(\underline{p})) \subset \text{Critical}_{D,F}(i)$$

In particular

$$V(D_i(\underline{p})) \subset \text{Crit}(D,F) := \bigcup_{i=1}^{k} \text{Critical}_{D,F}(i)$$

for every $i, \underline{p}$.

Proof. Let $\underline{x} \in V(D_i(\underline{p}))$. It is known that $V(D_i(\underline{p})) \subset \partial D_i(\underline{p})$, and so by Lemma 11.3, there must be some $P_1 \in \text{Planes}_{D,F}(i)$ such that $\underline{x} \in P_1$. If this is the only plane that $\underline{x}$ lies on in $\text{Planes}_{D,F}(i)$, then it is a simple exercise in elementary analysis (using the form of $D_i(\underline{p})$ from Corollary 11.4) to show that there exists an (m-1)-dimensional ball $B_{m-1}(\underline{x})$ centred on $\underline{x}$, such that $B_{m-1}(\underline{x}) \subset P_1 \cap D \subset D$. Therefore there is a line $L \subset B_{m-1}(\underline{x}) \subset D$ such that $\underline{x} = L(\lambda)$ for some $\lambda \in (0,1)$, contradicting the fact that $\underline{x}$ is a vertex. Hence, $\underline{x}$ must lie in at least two planes,

$P_1, P_2 \in \text{Planes}_{D,F}(i)$. Note that there is no m-ball $B_m(\underline{x}) \subset D$ since P_1 is on the boundary of $D_i(\underline{p})$.

More generally: Let $P(j) := P_1 \cap \cdots \cap P_j$, and let $d := \dim P(j)$. Then, the same line of reasoning can be used to show that if $\underline{x} \in P(j)$, then there is an d-dimensional ball $B_d(\underline{x})$ centred on $\underline{x}$, such that $B_d(\underline{x}) \subset P(j) \cap D \subset D$. It is likewise a simple exercise to show that there is no (d+1)-ball, $B_{d+1}(\underline{x}) \subset D$. Hence if $d > 0$, then as before, there will be a line $L \subset B_d(\underline{x}) \subset D$ such that $\underline{x} = L(\lambda)$ for some $\lambda \in (0, 1)$. Therefore, if $\underline{x} \in V(D_i(\underline{p}))$, then we need there to be some $P_1, \cdots, P_m \in \text{Planes}_{D,F}(i)$ such that $\underline{x} \in P_1 \cap \cdots \cap P_m =: P(m)$ where $\dim P(m) = 0$. Since these are hyperplanes, this implies that $\#P(m) = 1$. Hence $\underline{x} \in \text{Critical}_{D,F}(i)$, as required. □

Finally, we may use Lemma 11.6 to conclude the following:

Proposition 11.7. *Let D be a convex polytope and F a continuous, piecewise linear function where* $\text{Func}(F) = \{f_1, \cdots, f_k\}$. *Then*

$$\max_{\underline{x} \in D} F(\underline{x}) = \max_{\underline{x} \in \text{Crit}(D,F) \cap D} F(\underline{x}),$$

$$\min_{\underline{x} \in D} F(\underline{x}) = \min_{\underline{x} \in \text{Crit}(D,F) \cap D} F(\underline{x}).$$

Proof. By Corollary 11.5 and Lemma 11.6, we have

$$\begin{aligned} \max_{\underline{x} \in D} F(\underline{x}) &= \max_j \max_{\underline{x} \in V(D_{i_j}(\underline{p}_j))} F(\underline{x}) \\ &\leq \max_{\underline{x} \in \text{Crit}(D,F) \cap D} F(\underline{x}). \end{aligned}$$

We trivially have $\max_{\underline{x} \in D} F(\underline{x}) \geq \max_{\underline{x} \in \text{Crit}(D,F) \cap D} F(\underline{x})$. □

This allows us to create the following naive algorithm to find $\max_{\underline{x} \in D} F(\underline{x})$:

1. Construct a set of points S which contains $\text{Crit}(D, F)$.
2. Remove any points of S which do not lie in D (i.e. construct $S \cap D$).
3. Compute $F(\underline{x})$ for every $\underline{x} \in S \cap D$.

4. Return the largest value from the computed $\underline{x}$'s. This is $\max_{\underline{x}\in D} F(\underline{x})$ by Proposition 11.7.

11.1 A Simple Example

Unfortunately, even the most simple (non-trivial) examples have quite a few points that need to be tested, so we will not list them all: Let $D := \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1,\ -1 \leq y \leq -x\}$ and let

$$F(x, y) := \max\{2x + y, \min\{1, x/2 - 2y\}\}.$$

Then we have

$$\begin{aligned}\mathrm{Crit}(D, F) \subset &\{2x + y = 1 = x/2 - 2y\} \cup \{x = 0,\ 2x + y = 1\} \\ &\cup \{x = 0,\ 2x + y = x/2 - 2y\} \cup \{x = 0,\ 1 = x/2 - 2y\} \\ &\cup \{x = 1,\ 2x + y = 1\} \cup \cdots \cup \{x = 0,\ y = -1\} \\ &\cup \{x = 0,\ y = -x\} \cup \{x = 1,\ y = -1\} \cup \{x = 1,\ y = -x\}.\end{aligned}$$

All we have done is considered all of the different ways in which the functional values of F can equal each other, along with the different ways in which the boundary conditions can intersect themselves or intersect with the different functional values of F. We may now simplify these conditions (for example $2x + y = 1 = x/2 - 2y$ simplifies to $(x, y) = (2/3, -1/3)$) and remove any points that lie outside of D, as well as any duplicates. If we do this, we end up with

$$\mathrm{Crit}(D, F) \cap D \subset \{(2/3, -1/3), (0, 0), (0, -1/2), (1, -1), \cdots, (0, -1)\}.$$

After this, all that is left to do is evaluate $F(2/3, -1/3), F(0, 0)$ etc. The largest value of these will be $\max_D F(x, y)$ by Proposition 11.7.

Chapter 12

Minor Arcs Estimate

In this Chapter, we will combine all of the approaches we have been developing throughout this thesis to finally prove Proposition 5.3. In particular we aim to show that, provided $n \geq 39$, we have

$$S_{\mathfrak{m}} = O(P^{n-6-\delta})$$

for some $\delta > 0$. To achieve this, we will split the q sum of $S_{\mathfrak{m}}$ into square-free, cube-free square, 4th power-free cube, and 4th power-full parts (b_1, b_2, b_3, q_4 respectively), and further split these sums into $O(P^{\varepsilon})$ dyadic ranges. In particular, we will be focusing on the sum

$$D_P(R, t, \underline{R}) := \sum_{b_1=R_1}^{2R_1} \sum_{b_2=R_2}^{2R_2} \sum_{b_3=R_3}^{2R_3} \sum_{q_4=R_4}^{2R_4} {\sum_{\underline{a}}}^{*} \int_{t \ll |\underline{z}| \ll t} |S_{\underline{a}}(q, \underline{z})| \, d\underline{z},$$

where, $\underline{R} := (R_1, R_2, R_3)$, and

$$q = b_1 b_2 b_3 q_4, \quad R < q \leq 2R, \quad R_i < b_i \leq 2R_i, \, i \in \{1, 2, 3\}, \quad R_4 < q_4 \leq 2R_4 \tag{12.1}$$

(the latter is apparent from the definition of $D_P(R, t, \underline{R})$, but it will be helpful to be able to reference this later). From the definition of $S_{\mathfrak{m}}$, we need only consider $D_P(R, t, \underline{R})$ when

$$R \leq Q, \quad R_1 R_2 R_3 R_4 \asymp R, \quad 0 \leq t \leq (RQ^{1/2})^{-1}. \tag{12.2}$$

Likewise, we must also either have

$$R \geq P^{\Delta} \quad \text{or} \quad t \geq P^{-4+\Delta}. \tag{12.3}$$

Now, upon bounding $S_{\underline{a}}(q, \underline{z})$ trivially for $t \leq P^{-5}$, we see that

$$S_{\mathfrak{m}} \ll P^{\varepsilon} \max_{\substack{R,\underline{R},\underline{t} \\ (12.2),(12.3),\, t>P^{-5}}} D_P(R, t, \underline{R}) + O(P^{n-7}) \tag{12.4}$$

Our aim in this chapter is to show that $D_P(R, t, \underline{R}) \ll P^{n-6-\delta}$ for some $\delta > 0$, as this is sufficient to bound our minor arcs by $P^{n-6-\delta}$ by (12.4). Note that this is equivalent to proving that

$$\log_P(D_P(R, t, \underline{R})) := B_P(\phi, \tau, \underline{\phi}) \leq n - 6 - \delta \tag{12.5}$$

for some $\delta > 0$, and for P sufficiently large (so that the implied constant in (12.4) becomes negligible), where

$$\phi := \log_P(R), \quad \tau := \log_P(t), \quad \log_P(R_i) := \phi_i,\ i \in \{1, 2, 3, 4\}. \tag{12.6}$$

Finally, as mentioned in Chapter 5 we will choose

$$Q \asymp P^{3/2}$$

from this point onwards (this choice will be explained in Subsection 12.3.1). With this last bit of setup, we are now ready to start the process of bounding $S_{\mathfrak{m}}$. We will do this by applying the various bounds we have found in previous sections to $D_P(R, t, \underline{R})$ for different ranges of R and t. The process of covering all possible values of R and t will unfortunately be rather complicated. Throughout this section, we will use the following Lemma:

Lemma 12.1. *Let $q = b_1 b_2 \cdots b_k q_{k+1}$, where b_i is the ith power, (i+1)th powerfree part of q and let q_{k+1} be the (k+1)th power-full part of q. Then*

$$\sum_{\substack{b_i = R_i \\ i \in \{1, \cdots, k\}}}^{2R_i} \sum_{q_{k+1}=R_{k+1}}^{2R_{k+1}} b_1^{a_1} b_2^{a_2} \cdots b_k^{a_k} q_3^{a_{k+1}} \ll \prod_{i=1}^{k+1} R_i^{a_i + 1/i}$$

for every $a_1, \cdots, a_{k+1} \geq 0$.

The proof of this lemma is standard, and is similar to [3, Lemma 20] so we omit here. This Lemma enables us to get away with using slightly worse exponential sum bounds for the perfect square and cube-full parts of q (close inspection of the bounds found in Section 8 will show that our bounds in these cases are indeed worse). We have stated Lemma 12.1 in this level of generality because it will be useful for us when considering the singular series of the major arcs. We will spend the remainder of this section finding our final bounds for the minor arcs. We will find a total of five different bounds based on different combinations of van der Corput differencing, Weyl differencing, and Poisson summation.

12.1 Averaged van der Corput/Poisson

In this section, we will find a bound for $B_P(\phi, \tau, \underline{\phi}) := \log_P(D_P(R, t, \underline{R}))$ by combining the improved averaged van der Corput differencing process with Poisson summation. We will aim to show that $B_P(\phi, \tau, \underline{\phi}) \leq n - 6 - \delta$ for some $\delta > 0$, provided that n is sufficiently large. By Proposition 7.5, we have

$$D_P(R, t, \underline{R}) \ll_{\varepsilon, N} P^{-N} + \sum_{q, (12.1)} H^{-n/2+1} P^{n/2-1+\varepsilon} q((HP^2)^{-1} + t)^2 \times \Big(\max_{\underline{z}} \sum_{|\underline{h}| \ll H} |T_{\underline{h}}(q, \underline{z})| \Big)^{1/2}, \tag{12.7}$$

where $|\underline{z}| \asymp \max\{(HP^2)^{-1}, t\}$. By Proposition 10.2 we have

$$\sum_{\underline{h} \ll H} |T_{\underline{h}}(q, \underline{z})| \ll q^2 P^{n+\varepsilon} \Big\{ 1 + \sum_{\eta=0}^{n-1} \mathcal{Y}_\eta \Big\}$$

where

$$\mathcal{Y}_\eta(q,b_1,b_3,q_4,|\underline{z}|) := \frac{H^{n-\eta}}{q^{(n-\eta)/2}} b_1^{-1}\Big(V^{n-\eta} + b_1^{1/2}b_3^{1/3}q_4^{1/2}V^{n-\eta-1} + b_1^{1/2}b_3^{(n-\eta)/3}q_4^{(n-\eta)/2}\Big), \tag{12.8}$$

for $\eta \in \{0,\cdots,n-2\}$,

$$\mathcal{Y}_{n-1} := \frac{H}{q^{1/2}} b_1^{-1}(b_1^{1/2}V + b_1^{1/2}b_3^{1/3}q_4^{1/2}), \tag{12.9}$$

and

$$H(q) := \max\{P^{10/(n-2)+\varepsilon'}, P^{2/(n+2)+\varepsilon'}q^{6/(n+2)}\} \tag{12.10}$$

$$V(q,|\underline{z}|) := 1 + qP^{\varepsilon-1}\max\{1,\sqrt{HP^2|\underline{z}|}\}. \tag{12.11}$$

It is currently difficult to explain this choice of H, but we will justify this in Subsection 12.1.1. We note that $V(q,|\underline{z}|) \asymp V(q,t)$ in the range of $\underline{z}$ that we have. Hence (assuming

N is chosen sufficiently large):

$$\begin{aligned} D_P(R,t,\underline{R}) &\ll \sum_{q,(12.1)} H^{-n/2+1}P^{n-1+\varepsilon}q^2((HP^2)^{-1}+t)^2\times \\ &\qquad\qquad \Big(1+\sum_{\eta=0}^{n-1}\mathcal{Y}_\eta(q,b_1,b_3,q_4,t)\Big)^{1/2} \\ &\ll P^{n-1+\varepsilon}\sum_{\substack{b_i=R_i\\ i\in\{1,2,3\}}}^{2R_i}\sum_{q_4=R_4}^{2R_4} R^2H^{-n/2+1}((HP^2)^{-1}+t)^2\times \\ &\qquad\qquad (1+\mathcal{Y}_0+\cdots+\mathcal{Y}_{n-1})^{1/2} \qquad (12.12) \\ &\ll P^{n-1+\varepsilon}\mathcal{R}R_1^{1/2}R^2H^{-n/2+1}((HP^2)^{-1}+t)^2(1+\mathcal{Y}_0+\cdots+\mathcal{Y}_{n-1})^{1/2}, \end{aligned} \tag{12.13}$$

where $\mathcal{R} := R_1^{1/2}R_2^{1/2}R_3^{1/3}R_4^{1/4}$, $H = H(R)$, $V = V(R,t)$, and

$\mathcal{Y}_i = \mathcal{Y}_i(R,R_1,R_3,R_4,t)$ in (12.12)-(12.13). For the most part, we will continue to use H, V, and $\mathcal{Y}_i$ instead of $H(R)$, $V(R,t)$ and $\mathcal{Y}_i(R,R_1,R_3,R_4,t)$ to avoid making the algebra more complicated than it already is. The final assertion is by Lemma 12.1.

We will start by simplifying the right-most bracket:

Lemma 12.2. *For every R, R_1, R_3, R_4, t satisfying* (12.2)*, we have*

$$(1 + \mathcal{Y}_0 + \cdots + \mathcal{Y}_{n-1}) \ll (1 + \mathcal{Y}_0).$$

Proof. For this proof, we will introduce the following sequence:

$$\mathcal{Y}'_\eta := \frac{H^{n-\eta}}{R^{(n-\eta)/2}} R_1^{-1} \Big(R_1^{\eta/n} V^{n-\eta} + R_1^{1/2} {R_3}^{1/3-\eta/3n} R_4^{1/2-\eta/2n} V^{n-\eta-1+\eta/n} + R_1^{1/2} R_3^{(n-\eta)/3} R_4^{(n-\eta)/2} \Big).$$

We will prove that this sequence has the following three properties:

1. $\mathcal{Y}_\eta \ll \mathcal{Y}'_\eta$ for every $\eta \in \{0, \cdots, n-1\}$.
2. $\mathcal{Y}'_0 = \mathcal{Y}_0$, and $\mathcal{Y}'_n \asymp 1$.
3. $\sum_{\eta=0}^{n} \mathcal{Y}'_\eta$ is a sum of three geometric series.

Verifying these three facts is sufficient to complete the proof since properties 1 and 2 imply that $(1 + \mathcal{Y}_0 + \cdots + \mathcal{Y}_{n-1}) \ll (\mathcal{Y}'_0 + \cdots + \mathcal{Y}'_{n-1} + \mathcal{Y}'_n)$, property 3 implies that $(\mathcal{Y}'_0 + \cdots + \mathcal{Y}'_{n-1} + \mathcal{Y}'_n) \ll (\mathcal{Y}'_0 + \mathcal{Y}'_n)$, and property 2 implies that $(\mathcal{Y}'_0 + \mathcal{Y}'_n) = (1 + \mathcal{Y}_0)$.

For property 1, we note that the term outside of the bracket of $\mathcal{Y}'_\eta$ is equal to the analogous term in $\mathcal{Y}_\eta$. It therefore suffices to bound each term in the bracket of $\mathcal{Y}_\eta$ from above by a term in the bracket of $\mathcal{Y}'_\eta$: We clearly have $V^{n-\eta} \leq R_1^{\eta/n} V^{n-\eta}$ when $\eta \in \{1, \cdots, n-2\}$ and $R_1^{1/2} V \leq R_1^{(n-1)/n} V$ for every $n \geq 2$. The third term of $\mathcal{Y}_\eta$ and $\mathcal{Y}'_\eta$ coincide with each other for every $\eta \in \{0, \cdots, n-1\}$.

As for the middle term, $R_1^{1/2} R_3^{1/3} R_4^{1/2} V^{n-\eta-1} \leq R_1^{1/2} R_3^{1/3-\eta/3n} R_4^{1/2-\eta/2n} V^{n-\eta-1+\eta/n}$ if and only if $V \geq R_3^{1/3} R_4^{1/2}$. However, if $V < R_3^{1/3} R_4^{1/2}$, then $R_1^{1/2} R_3^{1/3} R_4^{1/2} V^{n-\eta-1} \leq R_1^{1/2} R_3^{(n-\eta)/3} R_4^{(n-\eta)/2}$, which is the third term of $\mathcal{Y}'_\eta$. Hence we have $\mathcal{Y}_\eta \ll \mathcal{Y}'_\eta$.

Property 2 is trivial so we will move to verifying property 3. Again, we will go term by term: Let

$$\mathcal{Y}'_{\eta,1} := \frac{H^{n-\eta}}{R^{(n-\eta)/2}} R_1^{-1} \cdot R_1^{\eta/n} V^{n-\eta}.$$

Then

$$\mathcal{Y}'_{\eta+1,1} = HR^{-1/2}R_1^{1/n}V^{-1}\mathcal{Y}'_{\eta,1}$$

If we similarly define $\mathcal{Y}'_{\eta,2}$ and $\mathcal{Y}'_{\eta,3}$ in the obvious way, then we see that

$$\mathcal{Y}'_{\eta+1,2} = HR^{-1/2}R_3^{-1/3n}R_4^{-1/2n}V^{-1+1/n}\mathcal{Y}'_{\eta,1}, \quad \mathcal{Y}'_{\eta+1,3} = HR^{-1/2}R_3^{-1/3n}R_4^{-1/2n}\mathcal{Y}'_{\eta,3}.$$

Hence we may represent $\sum \mathcal{Y}'_\eta$ as a sum of three geometric series, as required. This completes the proof. □

We may use Lemma 12.2 to conclude that

$$D_P(R,t,\underline{R}) \ll P^{n-1+\varepsilon}\mathcal{R}R_1^{1/2}R^2H^{-n/2+1}((HP^2)^{-1}+t)^2(1+\mathcal{Y}_0)^{1/2}. \tag{12.14}$$

We now aim to simplify this expression further by showing that $V^n \leq R^{1/2}R_3^{1/3}R_4^{1/2}V^{n-1}$, or equivalently that $V \leq R^{1/2}R_3^{1/3}R_4^{1/2}$. Doing this, will let us show the following:

Lemma 12.3. *Let $Q = P^{3/2}$ and let H and V be defined as above. If $n \geq 23$ then*

$$V \leq R^{1/2}.$$

In particular

$$\mathcal{Y}_0 \leq R^{(1-n)/2}R_1^{-1}H^n(V^{n-1} + R_3^{n/3-1/2}R_4^{(n-1)/2})$$

Proof. We will firstly prove that $V \leq R^{1/2}$. Recall that

$$V = V(R,t) = 1 + RP^{-1+\varepsilon}\max\{1, H(R)P^2t\}^{1/2}.$$

We clearly have $1 \leq R^{1/2}$.

When $V = RP^{-1+\varepsilon}$, we note that $R > P^{1-\varepsilon}$ otherwise $RP^{-1+\varepsilon} \leq 1$, and so V cannot be equal to $RP^{-1+\varepsilon}$. Furthermore we see that $R \leq Q = P^{3/2}$, or equivalently $P^{-1} \leq R^{-2/3}$. Hence, provided that ε is chosen small enough so that $P^\varepsilon \leq R^{1/6}$, then we also have $P^{-1+\varepsilon} < R^{-1/2}$. Since $R > P^{1-\varepsilon}$, $\varepsilon < 0.1$ would suffice for example.

Finally, we consider when $V = RP^{-1+\varepsilon}(H(R)P^2t)^{1/2}$. In this case, since $t \leq (RQ^{1/2})^{-1}$,

$$V \leq R^{1/2}Q^{-1/4}P^{\varepsilon}\max\{P^{5/(n-2)+\varepsilon}, P^{1/(n+2)}R^{3/(n+2)}\}.$$

But since $R \leq Q$, $P < Q$ (and $5/(n-2) > 4/(n+2)$), we have

$$V < R^{1/2}Q^{5/(n-2)+\varepsilon-1/4} < R^{1/2},$$

provided $n \geq 23$.

This concludes the proof that $V \leq R^{1/2}$. For the second statement of the lemma, we start by noting that $V^n \leq R^{1/2}V^{n-1}$, and so by (12.8) and (12.2):

$$\begin{aligned}\mathcal{Y}_0(R, R_1, R_3, R_4, t) :&= \frac{H^n}{R^{n/2}}R_1^{-1}\Big(V^n + R_1^{1/2}R_3^{1/3}R_4^{1/2}V^{n-1} + R_1^{1/2}R_3^{n/3}R_4^{n/2}\Big)\\ &\leq \frac{H^n}{R^{n/2}}R_1^{-1}\Big(V^n + R^{1/2}V^{n-1} + R^{1/2}R_3^{n/3-1/2}R_4^{(n-1)/2}\Big)\\ &\ll \frac{H^n}{R^{n/2}}R_1^{-1}\Big(R^{1/2}V^{n-1} + R^{1/2}R_3^{n/3-1/2}R_4^{(n-1)/2}\Big)\\ &= R^{(1-n)/2}R_1^{-1}H^n(V^{n-1} + R_3^{n/3-1/2}R_4^{(n-1)/2}).\end{aligned}$$

□

Hence, if we let

$$\mathcal{X}_1(R, R_3, R_4, t) = \mathcal{X}_1 := R^{(1-n)/2}H(R)^nV(R,t)^{n-1} \tag{12.15}$$

$$\mathcal{X}_2(R, R_3, R_4) = \mathcal{X}_2 := R^{(1-n)/2}R_3^{n/3-1/2}R_4^{(n-1)/2}H(R)^n \tag{12.16}$$

then we now may Lemma 12.3 and (12.14) to bound $D_P(R, t, \underline{R})$ as follows:

$$\begin{aligned}D_P(R, t, \underline{R}) &\ll P^{n-1+\varepsilon}\mathcal{R}R^2H^{-n/2+1}((HP^2)^{-1} + t)^2(R_1 + \mathcal{X}_1 + \mathcal{X}_2)^{1/2})\\ &\ll P^{n-1+\varepsilon}R^{5/2}H^{(2-n)/2}\max\{(HP^2)^{-1}, t\}^2\max\{R, \mathcal{X}_1, \mathcal{X}_2\}^{1/2}. \quad (12.17)\end{aligned}$$

Finally, note that $D_P(R, t, \underline{R}) \ll P^{n-6-\delta}$ for some $\delta > 0$ if $\log_P(D_P(R, t, \underline{R})) \leq n - 6 - \delta$ (provided P is chosen large enough) and so it is sensible to consider bounding $B_P(\phi, \tau, \underline{\phi}) := \log_P(D_P(R, t, \underline{R}))$. By (12.17) and upon letting $R := P^{\phi}$,

$R_i := P^{\phi_i}$, $t := P^{\tau}$, we have

$$\begin{aligned} B_P(\phi,\tau,\underline{\phi}) &\ll \log_P(P^{n-1+\varepsilon}R^{5/2}H^{(2-n)/2}\max\{(HP^2)^{-1},t\}^2\max\{R,\mathcal{X}_1,\mathcal{X}_2\}^{1/2}) \\ &= n-1+\varepsilon+\frac{5\phi}{2}+\frac{(2-n)}{2}\cdot\log_P(H)+2\max\{-2-\log_P(H),\tau\}+ \\ &\quad \frac{1}{2}\max\{\phi,\log_P(\mathcal{X}_1),\log_P(\mathcal{X}_2)\}+\log_P(C) \end{aligned} \tag{12.18}$$

where C is the implied constant in (12.17). If P is made to be sufficiently large, $\log_P(C)$ can be absorbed into ε. Hence (recalling (12.10) - (12.11), (12.15)-(12.16)), if we set

$$\hat{H}(\phi) := \max\{\frac{10}{n-2}+\varepsilon', \frac{2}{n+2}+\varepsilon'+\frac{6\phi}{n+2}\} \tag{12.19}$$

$$\hat{V}(\phi,\tau) := \max\{0,-1+\phi,\phi+\frac{\tau+\hat{H}(\phi)}{2}\} \tag{12.20}$$

$$\tau\_\mathrm{brac}(\phi,\tau) := \max\{-2-\hat{H}(\phi),\,\tau\} \tag{12.21}$$

$$\begin{aligned} \mathcal{X}\_\mathrm{brac}(\phi,\tau,\phi_3,\phi_4) := \max\{&\phi,\ \frac{(1-n)\phi}{2}+n\,\hat{H}(\phi)+(n-1)\,\hat{V}(\phi,\tau), \\ &\frac{(1-n)\phi}{2}+\Big(\frac{n}{3}-\frac{1}{2}\Big)\phi_3+\frac{(n-1)\phi_4}{2}+n\,\hat{H}(\phi)\} \end{aligned} \tag{12.22}$$

(for some small $\varepsilon'>0$ that we may choose freely) then (12.18) gives us the following:

Lemma 12.4. *Let n be fixed, and*

$$\begin{aligned} B_{AV/P}(\phi,\tau,\phi_3,\phi_4) := n-1+\frac{5\phi}{2}+\frac{(2-n)}{2}\hat{H}(\phi)+2\tau\_\mathrm{brac}(\phi,\tau) \\ +\frac{1}{2}\mathcal{X}\_\mathrm{brac}(\phi,\tau,\phi_3,\phi_4). \end{aligned}$$

Then $B_{AV/P}(\phi,\tau,\phi_3,\phi_4)$ is a continuous, piecewise linear function, and for every $\varepsilon>0$, there is a sufficiently large P such that

$$B_P(\phi,\tau,\underline{\phi}) \le B_{AV/P}(\phi,\tau,\phi_3,\phi_4)+\varepsilon,$$

for every $\phi\in[0,3/2]$, $\phi_i\in[0,\phi]$, $\phi_1+\phi_2+\phi_3+\phi_4=\phi$, $\tau\in[-5,-\phi-0.75]$.

The naming convention used is to make it easier to parse the algorithm's input. For example, τ_brac and $\hat{H}$ correspond to *Tau_bracket* and *H_Poisson* respectively in the algorithm's code.

12.1.1 The Limiting Case

In this subsection, we will briefly illustrate why we should expect the condition $n \geq 39$ to appear in Proposition 5.3 (or equivalently, why we should expect $D_P(R,t,\underline{R}) \ll P^{n-6-\delta}$ to be true for $n \geq 39$). In general, we expect the limiting condition on n to be determined by the so-called "generic case" for $(R,t,\underline{R})$, which is

$$R = Q = P^{3/2}, \;\; \tau = (RQ^{1/2})^{-1} = P^{-9/4}, \;\; R_1 = R = P^{3/2}, \;\; R_2 = R_3 = R_4 = 1.$$

This is the case where R is as large as possible and is square-free, and t is as large as possible. In this case, we expect the averaged van der Corput/Poisson bound to dominate over the other bounds since it is our main bound. We will therefore pinpoint which component of (12.17) dominates and then solve this part by hand. When we do this, we will see that the condition $n \geq 39$ arises naturally.

Firstly, it is easy to check via the definitions of H and V (12.10) - (12.11) that when $R \asymp P^{3/2}$, $t \asymp P^{-9/4}$, $R_1 = R$, $R_2 = R_3 = R_4 = 1$, we have

$$H = \max\{P^{10/(n-2)+\epsilon'}, P^{11/(n+2)+\epsilon'}\}, \tag{12.23}$$

$$\begin{aligned} V &= P^{1/2} \max\{1,\; P^{10/(n-2)-1/4+\epsilon'},\; P^{11/(n+2)-1/4+\epsilon'}\}^{1/2} \\ &= P^{3/8+\epsilon'/2} \max\{P^{5/(n-2)},\; P^{11/2(n+2)}\}. \end{aligned} \tag{12.24}$$

We could remove the "1" term from V since $P^{10/(n-2)-1/4+\epsilon'} > 1$ when $n \leq 42$, and $P^{11/(n+2)-1/4+\epsilon'} > 1$ when $n \geq 42$. It makes sense to consider the cases $n \leq 42$ and $n > 42$ separately so that we can simplify H and V further. We will just consider $n \leq 42$ here to avoid repetition since this section is not necessary for our arguments overall.

In particular, when $n \leq 42$, then by (12.23) and (12.24), we have

$$H = P^{10/(n-2)+\epsilon'}, \quad V = P^{3/8+5/(n-2)+\epsilon'/2}. \tag{12.25}$$

We aim to insert these values into the right-hand side of (12.17), but we will firstly

perform some simplifications. In particular, we note that

$$\max\{(HP^2)^{-1}, t\} = \max\{P^{-2-10/(n-2)-\epsilon'}, P^{-9/4}\} = P^{-9/4} \tag{12.26}$$

since $n \leq 42$. Similarly by (12.15) - (12.16), we see that $\mathcal{X}_1 > \mathcal{X}_2$ since $R_3 = R_4 = 1$ and $V > 1$. Hence

$$\begin{aligned}
\max\{R, \mathcal{X}_1, \mathcal{X}_2\} &= \max\{R, R^{(1-n)/2} \cdot P^{10n/(n-2)+n\epsilon'} \cdot P^{3(n-1)/8+5(n-1)/(n-2)+(n-1)\epsilon'/2}\} \\
&= \max\{P^{3/2}, P^{3(1-n)/4+10n/(n-2)+3(n-1)/8+5(n-1)/(n-2)+\epsilon'}\} \\
&= \max\{P^{3/2}, P^{3(1-n)/8+(15n-5)/(n-2)+\epsilon'}\} \\
&= P^{3/2}
\end{aligned} \tag{12.27}$$

provided that $n \geq 38.8111\cdots + \epsilon'$. In other words, as long as $n \geq 39$ and ϵ' is chosen small enough, we have $\max\{R, \mathcal{X}_1, \mathcal{X}_2\} = R = P^{3/2}$. Inserting (12.25) - (12.27) into (12.17) gives the following:

$$\begin{aligned}
D_P(R, t, \underline{R}) &\ll P^{n-1+\varepsilon} R^{5/2} H^{(2-n)/2} \max\{(HP^2)^{-1}, t\}^2 \max\{R, \mathcal{X}_1, \mathcal{X}_2\}^{1/2} \\
&= P^{n-1+\varepsilon} \cdot P^{15/4} \cdot P^{[(2-n)/2]\times[10/(n-2)+\epsilon']} \cdot P^{-9/2} \cdot P^{3/4} \\
&= P^{n-1+18/4-5-9/2+\epsilon-(n-2)\epsilon'/2} \\
&= P^{n-6-\delta(\epsilon,\epsilon')},
\end{aligned}$$

where $\delta > 0$ provided that ϵ is chosen sufficiently small with respect to ϵ' (and $n > 2$). This verifies the generic case for $39 \leq n \leq 42$. One can use the same procedure when $n < 39$, using $\max\{R, \mathcal{X}_1, \mathcal{X}_2\} = P^{3(1-n)/8+(15n-5)/(n-2)+\epsilon'}$ instead of $P^{3/2}$. If one does this, it can be shown that $D_P(R, t, \underline{R}) \ll P^{n-6-\delta}$ if only if $n \geq 39$, leading to a contradiction. A similar process can also be done when $n > 42$, and this will return $D_P(R, t, \underline{R}) \ll P^{n-6-\delta}$ as one would expect.

H was chosen specifically to ensure that the term

$$P^{n-1+\varepsilon} R^{5/2} H^{(2-n)/2} \max\{(HP^2)^{-1}, t\}^2 R^{1/2} \ll P^{n-6-\delta}$$

for any value of n (we specifically care about $n \geq 39$ of course). This is critical for

the Poisson bound to return anything useful as this is one of the two main "limiting" components of $D_P(R,t,\underline{R})$ – the other being

$$P^{n-1+\varepsilon}R^{5/2}H^{(2-n)/2}\max\{(HP^2)^{-1},t\}^2\mathcal{X}_1^{1/2}.$$

Whilst it is the $\mathcal{X}_1$ component which ultimately gives us the "limiting" condition on n (since the R component of $D_P(R,t,\underline{R})$ works for $n \geq 2$), we are required to choose H to be of at least certain size in order for the R component to give the bound $P^{n-6-\delta}$. This in turn makes our $\mathcal{X}_1$ bound worse since there are positive powers of H appearing in it, so in reality, we attain $n \geq 39$ by optimising H in such a way that both of these components simultaneously give the bound of $P^{n-6-\delta}$ in the "generic" case.

This section was included to highlight the two components of our main bound which give the limiting condition on n, to show where this condition comes from, and to explain why H was chosen in the way that it was in (12.10). It also serves to highlight the importance of using an algorithm to automate this process in this situation. In particular, even in the case where we have specified $(R,t,\underline{R})$ (and these values are "nice" in some way), the calculations are already quite complicated, and this is arguably the easiest case to consider.

12.2 Pointwise van der Corput/Poisson

Next, we will find a bound for $B_P(\phi,\tau,\underline{\phi})$ by combining the improved Pointwise van der Corput differencing process with Poisson summation. This time, we may assume $t \ll |\underline{z}| \ll t$. By Propositions 7.1 and 10.2, the fact that the $\mathcal{Y}_i$s are a geometric series, and Lemmas 12.2-12.3, (using the same values for $\mathcal{Y}, V, H$), we have:

$$\begin{aligned} D_P(R,t,\underline{R}) &\ll \sum_{q,(12.1)} \int_{t\ll|\underline{z}|\ll t} H(q)^{-n/2}P^{n/2}q\Big(\sum_{\underline{h}\ll H}|T_{\underline{h}}(q,\underline{z})|\Big)^{1/2}d\underline{z} \\ &\ll P^{n+\varepsilon}\sum_{q,(12.1)}\int_{t\ll|\underline{z}|\ll t} H(q)^{-n/2}q^2(1+\mathcal{Y}_0(q,b_1,q_3,|\underline{z}|))^{1/2}d\underline{z} \end{aligned}$$

$$\ll P^{n+\varepsilon} \sum_{q,(12.1)} t^2 H(R)^{-n/2} R^2 (1 + \mathcal{Y}_0(R, R_1, R_3, t))^{1/2}$$
$$\ll P^{n+\varepsilon} \mathcal{R} t^2 H(R)^{-n/2} R^2 (R_1 + \mathcal{X}_1 + \mathcal{X}_2)^{1/2}$$
$$\ll P^{n+\varepsilon} R^{5/2} t^2 H(R)^{-n/2} (R + \mathcal{X}_1 + \mathcal{X}_2)^{1/2}. \tag{12.28}$$

where the $\mathcal{X}_i$s are defined as in (12.15)-(12.16). Taking logs and recalling the definitions (12.19)-(12.22) gives us

$$B_P(\phi, \tau, \underline{\phi}) \leq n + \varepsilon + \frac{5\phi}{2} + 2\tau - \frac{n}{2} H + \frac{1}{2} \mathcal{X}\_\text{bracket} + \log_P(C),$$

where C is the implied constant in (12.28). Hence, we arrive at the following:

Lemma 12.5. *Let n be fixed, $\log_P D_P(R, t, \underline{R}) := B_P(\phi, \tau, \underline{\phi})$, and*

$$B_{PV/P}(\phi, \tau, \phi_3, \phi_4) := n + \frac{5\phi}{2} + 2\tau - \frac{n}{2} H + \frac{1}{2} \mathcal{X}\_\text{bracket}.$$

Then $B_{PV/P}(\phi, \tau, \phi_3, \phi_4)$ is a continuous, piecewise linear function, and for every $\varepsilon > 0$, there is a sufficiently large P such that

$$B_P(\phi, \tau, \underline{\phi}) \leq B_{PV/P}(\phi, \tau, \phi_3, \phi_4) + \varepsilon,$$

for every $\phi \in [0, 3/2]$, $\phi_i \in [0, \phi]$, $\phi_1 + \phi_2 + \phi_3 + \phi_4 = \phi$, $\tau \in [-5, -\phi - 0.75]$.

12.3 Averaged van der Corput/Weyl

We will now find a bound for $B_P(\phi, \tau, \underline{\phi})$ using the Averaged van der Corput differencing process discussed in Section 7, followed by one Weyl differencing step as in Section 6.

To keep the notation from getting out of hand, we will start using q before swapping this out for the b_1, b_2, b_3, q_4 notation when it becomes relevant. By Proposition 7.5 (upon choosing N to be sufficiently large), we have

$$D_P(R,t,\underline{R}) \ll_{\varepsilon,N} P^{-N} + \sum_{q,(12.1)} H^{-n/2+1}P^{n/2-1+\varepsilon}q((HP^2)^{-1}+t)^2\times \Big(\max_{t\ll|\underline{z}|\ll t}\sum_{|\underline{h}|\ll H}|T_{\underline{h}}(q,\underline{z})|\Big)^{1/2}, \quad (12.29)$$

We may now use Proposition 6.6 and (12.1) - (12.2) to bound $T_{\underline{h}}(q,\underline{z})$ as follows:

$$|T_{\underline{h}}(q,\underline{z})| \ll R^2P^{n+\varepsilon}\times \Big(P^{-2}+H^2R^2t^2+R^2P^{-4}+R^{-1}H^2\min\{1,\frac{1}{HtP^2}\}\Big)^{(n-\sigma_\infty(\underline{h})-2)/4}. \quad (12.30)$$

In this subsection, we will choose

$$H \asymp \max\{R^{1/6},(RtP^2)^{1/5}\}. \quad (12.31)$$

We will discuss the reason for this choice of H later, but for now, we note that $H=(RtP^2)^{1/5}$ when $t\geq(HP^2)^{-1}$, and $H=R^{1/6}$ when $t\leq(HP^2)^{-1}$ (this is easy to check using the definition of H). This is convenient for us since considering these two cases for t separately is natural due to the *min* bracket in (12.30).

Before we substitute (12.30) back into (12.29), we will simplify this expression significantly using the following Lemma:

Lemma 12.6. *Let $q\asymp R\leq Q$, $Q=P^{3/2}$, $|\underline{z}|\asymp t\leq(qQ^{1/2})^{-1}$, and $|\underline{h}|\ll H$, where H is defined as in* (12.31)*. Finally let $\sigma_\infty(\underline{h}):=s_\infty(F_{\underline{h}},G_{\underline{h}})$. Then*

$$T_{\underline{h}}(q,\underline{z}) \ll R^2P^{n+\varepsilon}\Big(R^{-1}H^2\min\{1,\frac{1}{HtP^2}\}\Big)^{(n-\sigma_\infty(\underline{h})-2)/4}.$$

Proof. Firstly we will assume that $t>(HP^2)^{-1}$. In this case the right-most term simplifies to $H/(RtP^2)$. Before we get into the proof that $H/(RtP^2)$ dominates all other terms, we will show the for our choice of H (see (12.31)), the following is true:

$$H \ll P^{1/4} \quad (12.32)$$

Indeed,

$$H \asymp (RtP^2)^{1/5} \ll Q^{-1/10}P^{2/5} \asymp P^{2/5-3/20} = P^{1/4}.$$

This will be useful to us as we attempt to show that $H/(RtP^2)$ dominates all other terms for every value of t and R. We now turn to proving this. Going from left to right in the bracket of (12.30), we firstly see that

$$P^{-2} \ll \frac{H}{RtP^2} \quad \Leftrightarrow \quad H \gg Rt.$$

But, we know that $t \leq (RQ^{1/2})^{-1}$, and so $Rt \ll 1$. We certainly have that $H \gg 1$, and so $H \gg Rt$ must be true. Next,

$$H^2R^2t^2 \ll \frac{H}{RtP^2} \quad \Leftrightarrow \quad HR^3t^3P^2 \ll 1.$$

Using the fact that $H \ll P^{1/4}$ by (12.32), and $Q \asymp P^{3/2}$ and $Rt \ll Q^{-1/2}$ by the assumptions in the Lemma, we see that

$$HR^3t^3P^2 \ll P^{1/4}Q^{-3/2}P^2 \asymp P^{9/4}(P^{-3/2})^{3/2} = 1,$$

as required. Finally,

$$R^2P^{-4} \ll \frac{H}{RtP^2} \quad \Leftrightarrow \quad H \gg R^3tP^{-2}.$$

This one has a few more steps. Recall that we are trying to show the dominance of the right term for every $\underline{t}$ and R. By our choice of H and the fact that $t \ll (RQ^{1/2})^{-1}$, $R \leq Q$, we have

$$\begin{aligned}
& R^3tP^{-2} \ll H = (RtP^2)^{1/5} \ \forall\ \underline{t}, R, \\
\Leftrightarrow \quad & R^{14/5}t^{4/5}P^{-12/5} \ll 1 \ \forall\ \underline{t}, R, \\
\Leftrightarrow \quad & \max\{R\}^7 \max\{t\}^2 P^{-6} \ll 1, \\
\Leftrightarrow \quad & Q^7(RQ^{-1/2})^{-2}P^{-6} \ll 1, \\
\Leftrightarrow \quad & Q^4P^{-6} \ll 1, \\
\Leftrightarrow \quad & Q \ll P^{3/2},
\end{aligned}$$

which is true. Hence, for our choices of H and Q, we have shown that $H^2R^{-1}\min\{1,(HtP^2)^{-1})\} = H/(RtP^2)$ dominates over all other terms in the expression for every $R \leq Q \asymp P^{3/2}$ and $(HP^2)^{-1} \leq t \leq (RQ^{1/2})^{-1}$.

A similar set of arguments can be used in the case that $t < (HP^2)^{-1}$. In this case, we have $H = R^{1/6}$, and

$$H^2R^{-1}\min\{1,(HtP^2)^{-1})\} = H^2R^{-1} = R^{-2/3}.$$

Again going from left to right in the bracket of (12.30):

$$P^{-2} \ll H^2R^{-1} = R^{-2/3} \quad \Leftrightarrow \quad P^2 \gg R^{2/3} \quad \Leftrightarrow \quad R \ll P^3,$$

which is true since $R \leq Q \asymp P^{3/2}$. Next,

$$H^2R^2t^2 \ll H^2R^{-1} \quad \Leftrightarrow \quad R(Rt)^2 \ll 1 \quad \Leftrightarrow \quad RQ^{-1} \ll 1 \quad \Leftrightarrow \quad R \ll Q,$$

which is again true by our assumptions from the Lemma. We used the fact that $Rt \leq Q^{-1/2}$ since $t \leq (RQ^{1/2})^{-1}$. Finally

$$R^2P^{-4} \ll H^2R^{-1} = R^{-2/3} \quad \Leftrightarrow \quad R^{8/3} \ll P^4 \quad \Leftrightarrow \quad R \ll P^{3/2}.$$

This is also true since $R \leq Q \asymp P^{3/2}$. Hence, we have shown that $H^2R^{-1}\min\{1,(HtP^2)^{-1})\} = H^2R^{-1}$ dominates over all other terms in the expression for every $R \leq Q \asymp P^{3/2}$ and $t \leq (HP^2)^{-1}$. This completes the proof of the lemma. □

We could now substitute the results from Lemma 12.6 into (12.29) directly, but the expression is rather complicated so we will instead just focus on the $\underline{h}$ sum inside of the integral for now. Our treatment of it will be analogous to the proof of the $\underline{h}$ sum bound in Section 10, but it will be a much simpler process this time around. The reason for our choice of H will also become apparent as we deal with this sum. We aim to show the following:

Lemma 12.7. *Let $q \asymp R \leq Q$, $Q = P^{3/2}$, $|\underline{z}| \asymp t \leq (qQ^{1/2})^{-1}$, and $|\underline{h}| \ll H$, where H is defined as in* (12.31)*. Then*

$$\sum_{|\underline{h}| \ll H} |T_{\underline{h}}(q, \underline{z})| \ll_n R^2 P^{n+\varepsilon} H.$$

In particular, we save a factor of H^n over the trivial bound.

Proof. We will again consider the cases when $t \geq (HP^2)^{-1}$ and $t \leq (HP^2)^{-1}$ separately. Starting with $t \geq (HP^2)^{-1}$ first: By Lemma 12.6, we have

$$\begin{aligned}
\sum_{|\underline{h}| \ll H} |T_{\underline{h}}(q, \underline{z})| &\ll R^2 P^{n+\varepsilon} \sum_{i=-1}^{n-1} \sum_{\substack{|\underline{h}| \ll H \\ \sigma_\infty(\underline{h}) = i}} \Big(\frac{H}{RtP^2}\Big)^{(n-i-2)/4} \\
&\ll R^2 P^{n+\varepsilon} \max_{-1 \leq i \leq n-1} \#\{|\underline{h}| \ll H \mid \sigma_\infty(\underline{h}) = i\} \Big(\frac{H}{RtP^2}\Big)^{(n-i-2)/4} \\
&\ll R^2 P^{n+\varepsilon} \max_{-1 \leq i \leq n-1} H^{n-i-1} \Big(\frac{H}{RtP^2}\Big)^{(n-i-2)/4} \qquad (12.33)
\end{aligned}$$

by Lemma 10.1. Recall that when $t \geq (HP^2)^{-1}$, we have $H \asymp (R|t|P^2)^{1/5}$. This value for H has been chosen specifically so that $H = (H/(RtP^2))^{-1/4}$ when $t > (HP^2)^{-1}$. The reason for doing this is so that the product within the *max* bracket in (12.33) will become H. Indeed, substituting this value for H into (12.33) gives

$$\begin{aligned}
\sum_{|\underline{h}| \ll H} |T_{\underline{h}}(q, \underline{z})| &\ll R^2 P^{n+\varepsilon} \max_{-1 \leq i \leq n-1} (RtP^2)^{(n-i-1)/5} \Big(\frac{1}{(RtP^2)^{4/5}}\Big)^{(n-i-2)/4} \\
&= R^2 P^{n+\varepsilon} \max_{-1 \leq i \leq n-1} (RtP^2)^{(n-i-1)/5} (RtP^2)^{-(n-i-2)/5} \\
&= R^2 P^{n+\varepsilon} (RtP^2)^{1/5} \\
&= R^2 P^{n+\varepsilon} H.
\end{aligned}$$

In theory, it would be nice if we could choose H to be even larger, so that we get something smaller than $R^2 P^{n+\varepsilon} H$. However, if one chooses H to be larger than this value, then Lemma 12.6 becomes false (in particular, the term $H^2 R^2 t^2$ dominates when $H > P^{1/4}$). This is therefore the optimal choice for H when $t > (HP^2)^{-1}$.

The argument in the case the $t \leq (HP^2)^{-1}$ is almost identical. Recall that when

$t \leq (HP^2)^{-1}$, we have $H \asymp R^{1/6}$. By Lemma 12.6, we have

$$\begin{aligned}\sum_{|\underline{h}|\ll H} |T_{\underline{h}}(q,\underline{z})| &\ll R^2P^{n+\varepsilon} \sum_{i=-1}^{n-1} \sum_{\substack{|\underline{h}|\ll H\\ \sigma_\infty(\underline{h})=i}} \left(H^2R^{-1}\right)^{(n-i-2)/4}\\ &\ll R^2P^{n+\varepsilon} \max_{-1\leq i\leq n-1} \#\{|\underline{h}| \ll H \mid \sigma_\infty(\underline{h}) = i\}\, R^{-(n-i-2)/6}\\ &\ll R^2P^{n+\varepsilon} \max_{-1\leq i\leq n-1} H^{n-i-1}R^{-(n-i-2)/6}\\ &\ll_n R^2P^{n+\varepsilon}R^{1/6}\\ &\ll R^2P^{n+\varepsilon}H\end{aligned}$$

by Lemma 10.1, and by the fact that when $t \leq (HP^2)^{-1}$, we have $H \asymp R^{1/6}$. This value for H has again been chosen specifically so that $H^{n-i-1}R^{-(n-i-1)/6} = 1$ for every i. when $t > (HP^2)^{-1}$. For the same reasons as before, we cannot choose H to be larger than this without causing other issues, and so this makes our choice of H in (12.31) optimal for our situation. □

Substituting the result of Lemma 12.7 back into (12.29) gives

$$D_P(R,t,\underline{R}) \ll P^{n-1+\varepsilon} \sum_{q,(12.1)} H^{-n/2+3/2}R^2((HP^2)^{-1}+t)^2$$

Finally, we split the R sum into its cube-free and cube-full components, and use Lemma 12.1 as follows:

$$\begin{aligned}D_P(R,t,\underline{R}) &\ll P^{n-1+\varepsilon} \sum_{b_1=R_1}^{2R_1} \sum_{b_2=R_2}^{2R_2} \sum_{b_3=R_3}^{2R_3} \sum_{q_4=R_4}^{2R_4} R^2H(R,t)^{(3-n)/2}((H(R,t)P^2)^{-1}+t)^2\\ &\ll P^{n-1+\varepsilon}R^3R_2^{-1/2}R_3^{-2/3}R_4^{-3/4}H(R,t)^{(3-n)/2}((H(R,t)P^2)^{-1}+t)^2\\ &\ll P^{n-1+\varepsilon}R^3R_3^{-2/3}R_4^{-3/4}H(R,t)^{(3-n)/2}((H(R,t)P^2)^{-1}+t)^2. \qquad (12.34)\end{aligned}$$

Therefore, upon setting $R := P^\phi$, $R_i := P^{\phi_i}$, $t := P^\tau$ and (recall (12.31))

$$\hat{H}\_\text{Weyl}(\phi,\tau) := \max\left\{\frac{\phi}{6}, \frac{2+\phi+\tau}{5}\right\}, \qquad (12.35)$$

$$\tau\_\text{brac}(\phi,\tau); = \max\{-2 - \hat{H}\_\text{Weyl}(\phi,\tau),\ \tau\}, \tag{12.36}$$

we have:

$$B_P(\phi,\tau,\underline{\phi}) \leq n - 1 + \varepsilon + 3\phi - \frac{2\phi_3}{3} - \frac{3\phi_4}{4} + \log_P(C) + \frac{(3-n)}{2}\hat{H}\_\text{Weyl}(\phi,\tau) + 2\tau\_\text{brac}(\phi,\tau),$$

where C is the implied constant in (12.34). Hence, if P is chosen to be sufficiently large, we may absorb $\log_P(C)$ into ε, giving us the following:

Lemma 12.8. *Let n be fixed, and*

$$B_{AV/W}(\phi,\tau,\phi_3,\phi_4) := n-1+3\phi-\frac{2\phi_3}{3}-\frac{3\phi_4}{4}+\frac{(3-n)}{2}\hat{H}\_\text{Weyl}(\phi,\tau)+2\tau\_\text{brac}(\phi,\tau).$$

Then $B_{AV/W}(\phi,\tau,\phi_3,\phi_4)$ is a continuous, piecewise linear function, and for every $\varepsilon > 0$, there is a sufficiently large P such that

$$B_P(\phi,\tau,\underline{\phi}) \leq B_{AV/W}(\phi,\tau,\phi_3,\phi_4) + \varepsilon,$$

for every $\phi \in [0,3/2]$, $\phi_i \in [0,\phi]$, $\phi_1+\phi_2+\phi_3+\phi_4 = \phi$, $\tau \in [-5,-\phi-0.75]$.

12.3.1 Explaining the Choice of Q

As an aside, we will briefly explain our choice of $Q \asymp P^{3/2}$, as promised in Chapter 5. We see in the proof of Lemma 12.6, that the optimal choice for Q is $P^{3/2}$. In particular, if we choose any other value for Q, then we cannot simplify the Weyl bound to such a large extent. We normally optimise our choice for Q based on our main bound, which in this case is the averaged van der Corput/Poisson bound. This value for Q turns out to be

$$Q \asymp P^{4(n+3)/3(n-2)},$$

which is the choice of Q that guarantees $HP^2|\underline{z}| \ll 1$ for every $\underline{z}$ (optimising our V term), where H and V are defined as in (12.10) - (12.11). In the range of n that we are considering, this value is largest when $n = 39$, giving us $Q \asymp P^{1.5135\cdots}$, which is

very close to the optimal choice for the van der Corput/Weyl bounds. In the end, the author chose $Q \asymp P^{3/2}$ because it is simpler and it makes the van der Corput/Weyl bounds significantly easier to work with. Most importantly, this choice does not cause any issues for our Poisson bounds, since it is "almost" optimal.

12.4 Pointwise van der Corput/Weyl

In this subsection, we will find a bound for $B_P(\phi, \tau, \underline{\phi})$ by using Pointwise van der Corput differencing, followed by one Weyl step. We start by applying Lemma 7.1 to $D_P(R, t, \underline{R})$:

$$D_P(R,t,\underline{R}) \ll \sum_{q,(12.1)} \int_{t\ll|\underline{z}|\ll t} H^{-n/2}P^{n/2}q\Big(\sum_{\underline{h}\ll H}|T_{\underline{h}}(q,\underline{z})|\Big)^{1/2}d\underline{z}.$$

Upon setting $H := \max\{q^{1/6}, (qtP^2)^{1/5}\}$ again, we may use Lemma 12.7 and Proposition 6.6) to conclude that

$$\begin{aligned} D_P(R,t,\underline{R}) &\ll P^{n+\varepsilon}\sum_{q,(12.1)}\int_{t\ll|\underline{z}|\ll t} H(q,t)^{(1-n)/2}q^2 d\underline{z} \\ &\ll P^{n+\varepsilon}R^3R_3^{-2/3}R_4^{-3/4}t^2H(R,t)^{(1-n)/2}. \end{aligned} \tag{12.37}$$

Hence upon recalling (12.35), we have

$$B_P(\phi,\tau,\underline{\phi}) \le n+\varepsilon+3\phi-\frac{2\phi_3}{3}-\frac{3\phi_4}{4}+2\tau+\log_P(C)+\frac{1-n}{2}\hat{H}\_\mathrm{Weyl}(\phi,\tau)$$

where C is the implied constant in (12.37). Therefore, if P is chosen to be sufficiently large, we may absorb $\log_P(C)$ into ε, giving us the following:

Lemma 12.9. *Let n be fixed, and*

$$B_{PV/W}(\phi,\tau,\phi_3,\phi_4) := n+3\phi+2\tau-\frac{2\phi_3}{3}-\frac{3\phi_4}{4}+\frac{1-n}{2}\hat{H}\_\mathrm{Weyl}(\phi,\tau).$$

Then $B_{PV/W}(\phi,\tau,\phi_3,\phi_4)$ is a continuous, piecewise linear function, and for every $\varepsilon > 0$, there is a sufficiently large P such that

$$B_P(\phi,\tau,\underline{\phi}) \le B_{PV/W}(\phi,\tau,\phi_3,\phi_4)+\varepsilon,$$

for every $\phi \in [0, 3/2]$, $\phi_i \in [0, \phi]$, $\phi_1 + \phi_2 + \phi_3 + \phi_4 = \phi$, $\tau \in [-5, -\phi - 0.75]$.

12.5 Weyl

In this subsection, we will find a bound for $B_P(\phi, \tau, \underline{\phi})$ by using Weyl differencing twice. We start by applying Proposition 6.1 to $D_P(R, t, \underline{R})$:

$$D_P(R,t,\underline{R}) \ll P^{n+\varepsilon} \sum_{q,(12.1)} \sideset{}{^*}\sum_{\underline{a}} \int_{t \ll |\underline{z}| \ll t} \Big(P^{-4} + q^2|\underline{z}|^2 + q^2P^{-6} + q^{-1}\min\{1, \frac{1}{|\underline{z}|P^3}\}\Big)^{(n-1)/16} d\underline{z}.$$

Firstly, it is easy to use (12.1)-(12.3) to check that

$$\max\{P^{-4}, q^2P^{-6}\} \leq q^{-1}\min\{1, (|\underline{z}|P^3)^{-1}.$$

Hence

$$\begin{aligned} D_P(R,t,\underline{R}) &\ll P^{n+\varepsilon} \sum_{q,(12.1)} \sideset{}{^*}\sum_{\underline{a}} \int_{t \ll |\underline{z}| \ll t} \Big(q^2|\underline{z}|^2 + q^{-1}\min\{1, \frac{1}{|\underline{z}|P^3}\}\Big)^{(n-1)/16} d\underline{z} \\ &\ll P^{n+\varepsilon} \sum_{q,(12.1)} q^2t^2\Big(q^2t^2 + q^{-1}\min\{1, \frac{1}{tP^3}\}\Big)^{(n-1)/16} \\ &\ll P^{n+\varepsilon} \sum_{q,(12.1)} q^2t^2\Big(q^2t^2 + q^{-1}\min\{1, \frac{1}{tP^3}\}\Big)^{(n-1)/16} \\ &\ll P^{n+\varepsilon} R^3R_3^{-2/3}t^2\Big(R^2t^2 + R^{-1}\min\{1, \frac{1}{tP^3}\}\Big)^{(n-1)/16} \end{aligned} \tag{12.38}$$

As usual, we are interested in $\log_P(D_P(R, t, \underline{R}))$ since this will be piecewise linear. The bound above gives

$$B_P(\phi,\tau,\underline{\phi}) \leq n + \varepsilon + 3\phi + 2\tau - \frac{2\phi_3}{3} - \frac{3\phi_4}{4} + \log_P(C) + \frac{n-1}{16}\max\Big\{2\phi + 2\tau,\ -\phi + \min\{0, -3 - \tau\}\Big\},$$

where $\log_P(C)$ is the implied constant in (12.38). Therefore, upon setting

$$\text{Weyl\_brac}(\phi, \tau) := \max\Big\{2\phi + 2\tau,\ -\phi + \min\{0, -3 - \tau\}\Big\} \tag{12.39}$$

we arrive at the following bound for B_P:

Lemma 12.10. *Let n be fixed, $\log_P D_P(R,t,\underline{R}) := B_P(\phi,\tau,\underline{\phi})$, and*

$$B_{Weyl}(\phi,\tau,\phi_3,\phi_4) := n+3\phi+2\tau-\frac{2\phi_3}{3}-\frac{3\phi_4}{4}+\frac{n-1}{16}\,\text{Weyl\_brac}(\phi,\tau).$$

Then $B_{Weyl}(\phi,\tau,\phi_3,\phi_4)$ is a continuous, piecewise linear function, and for every $\varepsilon>0$, there is a sufficiently large P such that

$$B_P(\phi,\tau,\underline{\phi}) \le B_{Weyl}(\phi,\tau,\phi_3,\phi_4)+\varepsilon,$$

for every $\phi\in[0,3/2]$, $\phi_i\in[0,\phi]$, $\phi_1+\phi_2+\phi_3+\phi_4=\phi$, $\tau\in[-5,-\phi-0.75]$.

12.6 Proof of Proposition 5.3

Recall that our ultimate goal is to show that

$$S_{\mathfrak{m}} \ll P^{n-6-\delta},$$

for some $\delta>0$, for every $n\ge 39$. This is equivalent to having

$$\log_P(S_{\mathfrak{m}}) < n-6.$$

We assume that ρ is chosen sufficiently small to facilitate average van der Corput differencing bounds. We may now use all of the previous subsections to bound $\log_P(S_{\mathfrak{m}})$ by a continuous, piecewise linear function in three variables: By (12.4), we have

$$\log_P(S_{\mathfrak{m}}) \le \log_P(c_1)+\varepsilon+\max_{\substack{\phi,\underline{\phi},\tau\\ (12.2),(12.3),\,\tau>P^{-5}}}\{B_P(\phi,\tau,\underline{\phi}),n-7\},$$

where c_1 is the implied constant. We clearly have that $\log_P(c_1)+\varepsilon+n-7\le n-6-\varepsilon$ for sufficiently large P, so we will assume that this is the case. Hence by Lemmas 12.4-12.10, we have

$$\log_P(S_{\mathfrak{m}}) \le \varepsilon+\max\Big\{\min_{(\phi,\tau,\phi_3,\phi_4)\in D_1\cup D_2}\Big\{B_{AV/P}(\phi,\tau,\phi_3,\phi_4),B_{PV/P}(\phi,\tau,\phi_3,\phi_4),$$
$$B_{AV/W}(\phi,\tau,\phi_3,\phi_4),\,B_{PV/W}(\phi,\tau,\phi_3,\phi_4),$$

$$B_{Weyl}(\phi,\tau,\phi_3,\phi_4)\Big\},n-6-2\varepsilon\Big\}, \quad (12.40)$$

where

$$D_1 := \{(\phi,\tau,\phi_3,\phi_4)\in\mathbb{R}^3 \,:\, \Delta\le\phi\le 3/2,\ 0\le\phi_3\le\phi,\ -5\le\tau\le-\phi-3/4\}$$

$$D_2 := \{(\phi,\tau,\phi_3,\phi_4)\in\mathbb{R}^3 \,:\, 0\le\phi\le\Delta,\ 0\le\phi_3\le\phi,\ -3+\Delta\le\tau\le-\phi-3/4\}.$$

Since D_1 and D_2 are convex polytopes and the function which we have bounded $\log_P(S_{\mathfrak{m}})$ is continuous and piecewise linear for every $n\in\mathbb{N}$. Each region on which this function is linear is a convex polytope. It is well known that extremum value of such a function must be taken at a vertex of one of these polytopes. Therefore, one may numerically compute the exact maxima in (12.40). We compute this maxima two different ways and check that both values coincide:

The first way is to use an inbuilt Min-Max function in Mathematica that compares the two bounds. This algorithm can be found in Appendix A. An executable version of code can also be found in the author's Github page [27]. The author has also verified this using an open source python based algorithm (this can also be found in [27]).

After taking $\varepsilon' = 0.0001$ (see (12.19)), $\Delta = 1/7 - 0.001$, both numerical verifications proves that

$$\log_P(S_{\mathfrak{m}}) \le n - 6.00185$$

for every $(\phi,\tau,\phi_3,\phi_4)\in D_1\cup D_2$, provided that $39\le n\le 48$. The limiting case is when $n=39$, $\phi=3/2$, $\tau=-2.25$, $\phi_2=\phi_3=\phi_4=0$. When $n\ge 49$, we may instead refer to Birch [1].

Chapter 13

Major Arcs

Finally, we will complete the proof of Theorems 3.1-3.2 by showing that

$$S_{\mathfrak{M}} = C_X P^{n-6} + O(P^{n-6-\delta})$$

where

$$S_{\mathfrak{M}} = \sum_{q \leq P^{\Delta}} \sideset{}{^*}\sum_{\underline{a}}^{q} \int_{|\underline{z}|<P^{-3+\Delta}} S(\underline{a}/q + \underline{z}) d\underline{z}$$

and C_X is a product of local densities. Let

$$\mathfrak{S}(R) := \sum_{q=1}^{R} q^{-n} \sideset{}{^*}\sum_{\underline{a}}^{q} S_{\underline{a},q}, \quad \mathfrak{J}(R) := \int_{|\underline{z}|<R} \int_{\mathbb{R}^n} \omega(\underline{x}) e(z_1 F(\underline{x}) + z_2 G(\underline{x}))\, d\underline{x} d\underline{z},$$

where

$$S_{\underline{a},q} := \sum_{\underline{x} \bmod q} e_q(a_1 F(\underline{x}) + a_2 G(\underline{x})),$$

and

$$\mathfrak{S} := \lim_{R\to\infty} \mathfrak{S}(R), \quad \mathfrak{J} = \lim_{R\to\infty} \mathfrak{J}(R),$$

if the limits exist. We will start by showing the following:

Lemma 13.1. *Assume that* $n - \sigma \geq 34$ *and that* $\mathfrak{S}$ *is absolutely convergent, satisfying*

$$\mathfrak{S}(R) = \mathfrak{S} + O_{\phi}(R^{-\phi})$$

for some $\phi > 0$. Then provided that we have $\Delta \in (0, 1/7)$,

$$S_{\mathfrak{M}} = \mathfrak{S}\mathfrak{J}P^{n-6} + O_{\phi}(P^{n-6-\delta}).$$

Following the proof found in [3], the first step towards proving this lemma is to show that

$$S(\underline{\alpha}) = q^{-n}P^n S_{\underline{a},q} I(\underline{z}P^3) + O(P^{n-1+2\Delta}) \tag{13.1}$$

where

$$I(\underline{t}) := \int_{\mathbb{R}^n} \omega(\underline{x}) e(t_1 F(\underline{x}) + t_2 G(\underline{x})) d\underline{x},$$

for $\underline{t} \in \mathbb{R}^2$. In order to achieve this, we need to be able to separate $S(\underline{\alpha})$'s dependence on $\underline{a}$ from its dependence on $\underline{z}$. Write $\underline{x} = \underline{u} + q\underline{v}$, where $\underline{u}$ runs over the complete set of residues modulo q and recall that $\underline{\alpha} = \underline{a}/q + \underline{z}$. Then

$$S(\underline{\alpha}) = \sum_{\underline{u} \bmod q} e_q(a_1 F(\underline{u}) + a_2 G(\underline{u})) \sum_{v \in \mathbb{Z}} \Phi_{\underline{u}}(\underline{v}), \tag{13.2}$$

where

$$\Phi_{\underline{u}}(\underline{v}) = \omega\Big(\frac{\underline{u} + q\underline{v}}{P}\Big) e(z_1 F(\underline{u} + q\underline{v}) + z_2 G(\underline{u} + q\underline{v})).$$

In order to have it so that $\underline{a}$ and $\underline{z}$ are independent from each other, we will replace our $\underline{v}$ sum with a crude integral estimate which has no dependence on $\underline{u}$. In particular, we can use the fact that $\Phi_{\underline{u}}(\underline{v} + \underline{x}) = \Phi_{\underline{u}}(\underline{v}) + O(\max_{y \in [0,1]^n} |\nabla \Phi_{\underline{u}}(\underline{v} + \underline{y})|)$ for any $\underline{x} \in [0,1]^n$, to conclude the following:

$$\begin{aligned}\Big| \int_{\mathbb{R}^n} \Phi_{\underline{u}}(\underline{v}) d\underline{v} - \sum_{\underline{v} \in \mathbb{Z}^n} \Phi_{\underline{u}}(\underline{v}) \Big| &\leq \mathrm{meas}(\mathcal{S}) \max_{\hat{\underline{v}} \in \mathcal{S} \cap \mathbb{Z}} \Big| \int_{\hat{\underline{v}} + [0,1]^n} \Phi_{\underline{u}}(\underline{v}) d\underline{v} - \Phi_{\underline{u}}(\hat{\underline{v}}) \Big| \\ &\ll \mathrm{meas}(\mathcal{S}) \max_{\hat{\underline{v}} \in \mathcal{S} \cap \mathbb{Z}} \max_{y \in [0,1]^n} |\nabla \Phi_{\underline{u}}(\hat{\underline{v}} + \underline{y})|.\end{aligned}$$

We note that

$$\max_{\hat{\underline{v}} \in \mathcal{S} \cap \mathbb{Z}} \max_{y \in [0,1]^n} |\nabla \Phi_{\underline{u}}(\hat{\underline{v}} + \underline{y})| = \max_{\hat{\underline{v}} \in \mathcal{S}} |\nabla \Phi_{\underline{u}}(\hat{\underline{v}})|,$$

and so by the Leibniz rule we have

$$\left| \int_{\mathbb{R}^n} \Phi_{\underline{u}}(\underline{v}) d\underline{v} - \sum_{\underline{v} \in \mathbb{Z}^n} \Phi_{\underline{u}}(\underline{v}) \right| \ll P^n q^{-n} (q/P + q|\underline{z}|P^2)$$
$$= P^{n-1} q^{1-n} + |\underline{z}| P^{n+2} q^{1-n}$$

since $\mathcal{S}$ is an n-dimensional cube with sides of order $1 + P/q \leq 2P/q$. Hence, on setting

$P\underline{x} = \underline{u} + q\underline{v}$, we arrive at the following expression for $\sum_{\underline{v}} \Phi_{\underline{u}}(\underline{v})$:

$$\sum_{\underline{v} \in \mathbb{Z}^n} \Phi_{\underline{u}}(\underline{v}) = \frac{P^n}{q^n} \int_{\mathbb{R}^n} \omega(\underline{x}) e(z_1 P^3 F(\underline{x}) + z_2 P^3 G(\underline{x})) d\underline{x} + O(P^{n-1} q^{1-n} + |\underline{z}| P^{n+2} q^{1-n}).$$

We can therefore conclude that

$$S(\underline{\alpha}) = P^n q^{-n} S_{\underline{a},q} I(\underline{z} P^3) + O(P^{n-1} q + |\underline{z}| P^{n+2} q) \tag{13.3}$$

by (13.2). Since $|\underline{z}| \leq P^{-3+\Delta}$ and $q \leq P^{\Delta}$, we can now conclude that (13.1) is indeed true. Furthermore, by substituting (13.1) into $S_{\mathfrak{M}}$ and – for the error term – noting that the major arcs have measure $O(P^{-6+5\Delta})$ ($P^{-6+2\Delta}$ from the integrals, $P^{3\Delta}$ from the sums), we conclude that

$$S_{\mathfrak{M}} = P^{n-6} \mathfrak{S}(P^{\Delta}) \mathfrak{J}(P^{\Delta}) + O(P^{n-7+7\Delta}). \tag{13.4}$$

Since we have assumed $\mathfrak{S}(R) = \mathfrak{S} + O_{\phi}(R^{-\phi})$ for some $\phi > 0$, we can replace $\mathfrak{S}(P^{\Delta})$ with $\mathfrak{S}$ leading us to

$$S_{\mathfrak{M}} = P^{n-6} \mathfrak{S} \mathfrak{J}(P^{\Delta}) + O_{\phi}(P^{n-7+7\Delta} + P^{n-6-\Delta\phi}). \tag{13.5}$$

We will prove that this assumption is true in the next section. We now aim to show that we can replace $\mathfrak{J}(P^{\Delta})$ with $\mathfrak{J}$. In order to do this, we need $\mathfrak{J}$ to exist, and $|\mathfrak{J} - \mathfrak{J}(P^{\Delta})|$ to be sufficiently small. Now, it is easy to see that

$$\mathfrak{J} - \mathfrak{J}(R) = \int_{|\underline{t}| \geq R} I(\underline{t}) d\underline{t},$$

and so this motivates us to find a bound for the size of $I(\underline{t})$. We will show the

following:

Lemma 13.2. *Let*

$$\sigma = \max\{\dim \mathrm{Sing}_{\mathbb{C}}(X_F), \dim \mathrm{Sing}_{\mathbb{C}}(X_G), \dim \mathrm{Sing}_{\mathbb{C}}(X_F, X_G)\}.$$

Then

$$I(\underline{t}) \ll \min\{1, |\underline{t}|^{\sigma+1-n/16+\varepsilon}\}.$$

Proof. We will again follow the same procedure as in [3]. $I(\underline{t}) \ll 1$ is trivial since $|I(\underline{t})| \leq \mathrm{meas}(\mathcal{S})$ for every $\underline{t}$. For the second estimate, we can assume $|\underline{t}| > 1$. Then on taking $\underline{a} = 0$, $q = 1$ in (13.3) we get

$$S(\underline{\alpha}) = P^n O(|\underline{\alpha}|P^3) + O((|\underline{\alpha}|P^3 + 1)P^{n-1})$$

for any $P \geq 1$. Likewise, for $|\underline{\alpha}| < P^{-1}$, we can also use Proposition 6.1 with $\underline{a} = \underline{0}$, $q = 1$, to conclude that

$$S(\underline{\alpha}) \ll P^{n+\varepsilon}(|\underline{\alpha}|P^3)^{(\sigma+1-n)/16}.$$

Hence for such α, we may set $\underline{t} = \underline{\alpha}P^3$ and combine these estimates to get

$$I(\underline{t}) \ll |\underline{t}|^{(\sigma+1-n)/16}P^{\varepsilon} + |\underline{t}|P^{-1}$$

when $1 < |\underline{t}| < P^2$. Finally, we note that this is true for every $P \geq 1$ and $I(\underline{t})$ does not depend on P at all. Hence we can choose $P = |\underline{t}|^{(16+n-\sigma-1)/16}$ to reach our second estimate of $I(\underline{t})$. □

We can now use Lemma 13.2 to conclude that

$$\begin{aligned}\mathfrak{J} - \mathfrak{J}(R) = \int_{|\underline{t}|\geq R} I(\underline{t})d\underline{t} &\ll \int_R^{\infty}\int_R^{\infty} \min\{1, |\underline{t}|^{(\sigma+1-n)/16+\varepsilon}\}d\underline{t} \\ &\ll R^{(33+\sigma-n)/16+\varepsilon}.\end{aligned}$$

For $n - \sigma \geq 34$, this shows that $\mathfrak{J}$ is absolutely convergent. Finally, replacing $\mathfrak{J}(P^{\Delta})$

by $\mathfrak{J}$ in (13.5) gives us

$$S_{\mathfrak{M}} = \mathfrak{S}\mathfrak{J}P^{n-6} + O_\phi(P^{n-7+7\Delta} + P^{n-6-\Delta\phi} + P^{n-6-\Delta/16+\varepsilon})$$

which is permissible for Lemma 13.1 provided that $\Delta \in (0, 1/7)$, $\phi > 0$, and $\varepsilon > 0$ is taken to be sufficiently small.

13.1 Convergence of the singular series

Finally we turn to the issue of showing that the singular series

$$\mathfrak{S} := \sum_{q=1}^{\infty} q^{-n} \sideset{}{^*}\sum_{\underline{a}} S_{\underline{a},q}$$

converges absolutely, and obeys the assumption made in Lemma 13.1. In particular, we will show the following:

Theorem 13.3. *Assume $n-\sigma \geq 35$. Then $\mathfrak{S}$ is absolutely convergent. Furthermore, there is some $\phi > 0$ such that*

$$\mathfrak{S}(R) = \mathfrak{S} + O_\phi(R^{-\phi}).$$

To see that $\mathfrak{S}$ converges for $n-\sigma \geq 35$, we will again adopt the approach of Browning and Heath-Brown in [3]. We start by noting that

$$\mathfrak{S} = q^{-n} \sideset{}{^*}\sum_{\underline{a}}^{q} S_{\underline{a},q}$$

is a multiplicative function of q, and so it follows that $\mathfrak{S}$ is absolutely convergent if and only if $\prod_p(1 + \sum_{k=1}^{\infty} a_p(k))$ is, where

$$a_p(k) := p^{-kn} \sideset{}{^*}\sum_{\underline{a}}^{p^k} |S_{\underline{a},p^k}|.$$

But by taking logs, this is equivalent to $\sum_p \sum_{k=1}^{\infty} a_p(k)$ converging. Now by Proposition 6.1 with $\underline{a} = \underline{0}$, $q = p^k$, $|\underline{z}| < P^{-3+\Delta}$, $\omega = \chi$, we have that

$$a_p(k) \ll p^{k(2+(\sigma+1)/16-n/16)+\varepsilon} \tag{13.6}$$

for any $k \geq 1$, and so this enables us to establish that $\mathfrak{S}$ converges absolutely provided that $n - \sigma \geq 50$. We can use (13.6) far more effectively than this if we are more careful: We will assume that $n - \sigma \geq 35$ from now on. Then by (13.6), we have

$$\sum_p \sum_{k \geq 16} a_p(k) \ll \sum_p p^{33+\sigma-n+\varepsilon} < \sum_{m=1}^{\infty} m^{-2+\varepsilon} \ll 1,$$

assuming $\varepsilon > 0$ is sufficiently small. We now need to show that $\sum_p \sum_{1 \leq k \leq 15}$ also converges. For $2 \leq k \leq 15$, we will use [3, Lemma 25]. This shows that

$$S_{\underline{a},p^k} \ll_k p^{(k-1)n+s_p(a_1F+a_2G)+1}.$$

Hence

$$\sum_p \sum_{k=2}^{15} a_p(k) \ll \sum_p \sum_{k=2}^{15} p^{k(2-n)} p^{(k-1)n+s_p(a_1F+a_2G)+1} = \sum_p \sum_{k=2}^{15} p^{2k+1-n+s_p(a_1F+a_2G)}.$$

But by Proposition 4.4, we have $s_p(a_1F + a_2G) \leq s'_p(F, G) + 1$. Furthermore since F and G are fixed, $s'_p(F, G) = \sigma$ for all but finitely many primes, and so by increasing the size of the implicit multiplicative constant if necessary, we have that

$$\sum_p \sum_{k=2}^{15} p^{2k+2-n+\sigma} \ll \sum_p p^{32-n+\sigma} \ll 1,$$

since we have assumed $n - \sigma \geq 35$.

All that is left to check is $k = 1$. By Lemma 7 in [3], we have

$$\sum_p a_p(1) \ll \sum_p p^{2-n/2+(s_p(a_1F+a_2G)+1)/2} \ll \sum_p p^{3-n/2+\sigma/2} \ll 1.$$

This enables us to establish Theorem 13.3. Finally, we will follow the approach used in [21] to prove that there exists some $\phi > 0$ such that

$$\mathfrak{S}(R) = \mathfrak{S} + O_\phi(R^{-\phi}).$$

We will continue to work under the assumption that $n - \sigma \geq 35$. Firstly let

$$S_q := \sum_{\underline{a}}^{q\,*} \sum_{\underline{x}}^{q} e_q(a_1F(\underline{x}) + a_2G(\underline{x})).$$

Then, we have

$$|\mathfrak{S} - \mathfrak{S}(R)| \leq \sum_{q \geq R} q^{-n} |S_q|. \tag{13.7}$$

We will split q into several of its multiplicative components and bound each component separately. Let

$$b_i := \prod_{p^i || q} p^i, \quad q_i := \prod_{\substack{p^e || q \\ e \geq i}} p^e.$$

Then $q = q_k \prod_{i=1}^{k-1} b_i$ for every k (e.g. $q = b_1 b_2 q_3$). Recall that by Lemma 12.1, we have the following for any $R_1, \cdots, R_k > 0$:

$$\sum_{\substack{b_1 \sim R_1, \cdots, b_{k-1} \sim R_{k-1} \\ q_k \sim R_k}} 1 \ll \prod_{i=1}^{k} R_i^{1/i}. \tag{13.8}$$

We will use $k = 16$. Now

$$|S_q| \leq |S_{q_{16}}| \prod_{i=1}^{15} |S_{b_i}|.$$

We will bound each of these in turn:

$$|S_{q_{16}}| \ll q_{16}^{(15n+\sigma+1)/16+\varepsilon}$$

by Proposition 6.1. For $b_3, \cdots, b_{15}$, we split b_k into prime powers and use Lemma 25 from [3]:

$$|S_{p^k}| \ll \sideset{}{^*}\sum_{\underline{a}}^{p^k} p^{(k-1)n + s_p(a_1 F + a_2 G) + 1} \ll p^{(k-1)n+\sigma+2+2k}$$

for $p \gg 1$. Hence for $k \in \{3, \cdots, 15\}$,

$$|S_{b_k}| \ll b_k^{2+((k-1)n+\sigma+2)/k}$$

Finally for b_1, b_2, we use Lemma 7 from [3]. By following the same argument as for $S_{b_3}, \cdots, S_{b_15}$, we get

$$|S_{b_k}| \ll b_k^{2+(n+\sigma+2)/2},$$

for $k \in \{1, 2\}$. Hence

$$|S_q| \ll q^{2+\varepsilon} (b_1 b_2)^{(n+\sigma+2)/2} b_3^{(2n+\sigma+2)/3} \cdots b_{15}^{(14n+\sigma+2)/15},$$

or equivalently

$$|S_q| \ll \frac{q^{2+n+\varepsilon}}{(b_1b_2)^{(m-1)/2}b_3^{(m-1)/3}\cdots b_{15}^{(m-1)/15}q_{16}^{m/16}},$$

where $m = n - \sigma - 1$. Therefore, by (13.7), we have

$$\begin{aligned}|\mathfrak{S} - \mathfrak{S}(R)| &\ll \sum_{b_1\cdots b_{15}q_{16}\geq R} (b_1b_2)^{2+\varepsilon-(m-1)/2}b_3^{2+\varepsilon-(m-1)/3}\cdots b_{15}^{2+\varepsilon-(m-1)/15}q_{16}^{2+\varepsilon-m/16}\\ &\ll \sum_{b_1\cdots b_{15}q_{16}\geq R} (b_1b_2)^{(5+\varepsilon-m)/2}b_3^{(7+\varepsilon-m)/3}\cdots b_{15}^{(31+\varepsilon-m)/15}q_{16}^{(32+\varepsilon-m)/16}.\end{aligned}$$

When $m \geq 34$, we clearly have

$$\begin{aligned}|\mathfrak{S} - \mathfrak{S}(R)| &\ll \sum_{b_1\cdots b_{15}q_{16}\geq R} (b_1b_2)^{-29/2+\varepsilon}b_3^{-27/3+\varepsilon}\cdots b_{15}^{-3/15+\varepsilon}q_{16}^{-2/16+\varepsilon}\\ &\ll R^{-1/16+2\varepsilon}\sum_{b_1\cdots b_{15}q_{16}\geq R} (b_1b_2)^{-1-\varepsilon}b_3^{-1/3-\varepsilon}\cdots b_{15}^{-1/15-\varepsilon}q_{16}^{-1/16-\varepsilon}\\ &< R^{-1/16+2\varepsilon}\sum_{b_1,\cdots,b_{15},q_{16}=1}^{\infty} (b_1b_2)^{-1-\varepsilon}b_3^{-1/3-\varepsilon}\cdots b_{15}^{-1/15-\varepsilon}q_{16}^{-1/16-\varepsilon}\end{aligned}$$

and this sum converges by (13.8). Hence, we conclude that

$$\mathfrak{S} = \mathfrak{S}(R) + O(R^{-\phi}),$$

where $\phi = 1/16 - \varepsilon$, provided that $n - \sigma \geq 35$.

13.2 Proving that $\mathfrak{J} > 0$

In Section 13.1, we proved that $\mathfrak{J}$ was absolutely convergent provided that $n-\sigma \geq 34$, where

$$\mathfrak{J}(R) := \int_{|\underline{z}|<R}\int_{\mathbb{R}^n} \omega(\underline{x})e(z_1F(\underline{x}) + z_2G(\underline{x}))\, d\underline{x}d\underline{z}, \quad \mathfrak{J} := \lim_{R\to\infty}\mathfrak{J}(R). \tag{13.9}$$

In order to complete the proof of Theorem 3.2, we must show that $C_X = \mathfrak{S}\mathfrak{J} > 0$ in the case when $\sigma = -1$. In this section, we will focus on proving that the singular integral, $\mathfrak{J}$ is greater than 0. The argument to show this is very standard, and so we will simply follow the relevant parts of similar proofs used by Browning, Dietmann, and Heath-Brown in [4] and Davenport in [9].

It suffices to show that $\mathfrak{J}(R) \gg 1$ for R sufficiently large in order to prove that $\mathfrak{J} > 0$. The most natural way to do this is to explicitly integrate $\mathfrak{J}(R)$ and see what this gives us. Indeed, if we let $\underline{x} = \underline{x}_0 + \underline{y}$, then by permuting integrals and considering the odd and even decomposition of $e(\underline{x})$, we have the following:

$$\begin{aligned}\mathfrak{J}(R) &= \int_{-R}^{R}\int_{-R}^{R}\int_{\mathbb{R}^n} \omega(\underline{x})e(z_1F(\underline{x}) + z_2G(\underline{x}))\, d\underline{x}d\underline{z} \\ &= \int_{\mathbb{R}^n} \omega(\underline{x})\frac{\sin\left(2\pi RF(\underline{x})\right)\sin\left(2\pi RG(\underline{x})\right)}{\pi^2F(\underline{x})G(\underline{x})}\, d\underline{x} \\ &= \int_{\mathbb{R}^n} \gamma(\rho^{-1}\underline{y})\frac{\sin\left(2\pi RF(\underline{x}_0+\underline{y})\right)\sin\left(2\pi RG(\underline{x}_0+\underline{y})\right)}{\pi^2F(\underline{x}_0+\underline{y})G(\underline{x}_0+\underline{y})}\, d\underline{y}, \end{aligned} \tag{13.10}$$

where ω and γ are defined as in (3.6) and (3.5) respectively. As in previous chapters, $\underline{x}_0$ is the non-singular solution to $F(\underline{x}) = G(\underline{x}) = 0$ that we chose to centre our weight function ω on, and $\rho \in (0,1)$ is a constant which we may choose freely (some constraints have already been placed on ρ in Chapter 5). From here, the hope is that we can show that $F(\underline{x}_0 + \underline{y})$ and $G(\underline{x}_0 + \underline{y})$ are "relatively close" to zero for every $\underline{y} \in \rho\,\mathrm{Supp}(\gamma)$ because if this is true, then this makes it likely that the expression within the integral of (13.10) will be bounded away from zero.

For convenience, let $a_i := \partial F/\partial x_i(\underline{x}_0)$, and $b_i := \partial G/\partial x_i(\underline{x}_0)$ for $i \in \{1, \cdots, n\}$ and note that since we have assumed that $\mathrm{Rank}(\nabla F(\underline{x}_0), \nabla G(\underline{x}_0)) = 2$, there must be some i, j such that $a_ib_j - a_jb_i \neq 0$. We will assume $i = 1$, $j = 2$ without loss of generality. We have taken $\underline{x}_0$ to be a fixed point, so we may expand $F(\underline{x}_0 + \underline{y})$ and $F(\underline{x}_0 + \underline{y})$ as follows:

$$\begin{aligned} u_1 = u_1(\underline{y}) := F(\underline{x}_0 + \underline{y}) &= F(\underline{x}_0) + \underline{y}\cdot\nabla F(\underline{x}_0) + P_2(\underline{y}) + P_3(\underline{y}) \\ &= a_1y_1 + \cdots + a_ny_n + P_2(\underline{y}) + P_3(\underline{y}), \end{aligned}$$

$$u_2 = u_2(\underline{y}) := G(\underline{x}_0 + \underline{y}) = b_1 y_1 + \cdots + b_n y_n + Q_2(\underline{y}) + Q_3(\underline{y}),$$

where P_i, Q_i are of degree i (also recall that $F(\underline{x}_0) = G(\underline{x}_0) = 0$). From here, we will directly follow the second half of the proof laid out in [9, Chapter 16] as the argument used generalises trivially to two cubics. We see that $u_1, u_2 \ll \rho$ for every $|\underline{y}| \leq \rho$, and so we may use the inverse function theorem to represent y_1 and y_2 as a power series in $u_1, u_2, y_3, \cdots, y_n$, provided that ρ is chosen to be small enough. In particular, we have

$$\begin{aligned} y_1 &= u_1 - a_3 y_3 - \cdots - a_n y_n + \hat{P}_1(u_1, y_3, \cdots, y_n) \\ y_2 &= u_2 - b_3 y_3 - \cdots - b_n y_n + \hat{P}_2(u_2, y_3, \cdots, y_n), \end{aligned}$$

where $\hat{P}_i$ are multiple power series beginning with terms of at least degree 2. We may now take derivatives to see that for $i \in \{1, 2\}$,

$$\frac{\partial y_i}{\partial u_i} = 1 + \frac{\partial \hat{P}_i(u_i, y_3, \cdots, y_n)}{\partial u_i}$$

which implies that

$$\frac{1}{2} < 1 + \frac{\partial \hat{P}_i(u_i, y_3, \cdots, y_n)}{\partial u_i} < \frac{3}{2} \tag{13.11}$$

for ρ sufficiently small, since $|y_3|, \cdots, |y_n| < \rho$ and $|u_1|, |u_2| \ll \rho$. We may now use this to perform a change of variables from y_1, y_2 to u_1, u_2 respectively in (13.10) to get

$$\mathfrak{J}(R) = \int_{|u_1| \ll \rho} \int_{|u_2| \ll \rho} \frac{\sin(2\pi R u_1) \sin(2\pi R u_2)}{\pi^2 u_1 u_2} V(\underline{u})\, d\underline{u}, \tag{13.12}$$

where

$$V(\underline{u}) := \int_{B'} 1 + \frac{\partial \hat{P}_i(u_i, y_3, \cdots, y_n)}{\partial u_i} dy_3 \cdots dy_n,$$

B' is the $(n-2)$ dimensional cube $|y_3|, \cdots, |y_n| < \rho$, which implies that

$$V(\underline{u}) \geq \rho^{n-2}/2 > 0 \tag{13.13}$$

by (13.11). Furthermore

$$\lim_{u_1 \to 0} \lim_{u_2 \to 0} \frac{\sin(2\pi R u_1)\sin(2\pi R u_2)}{\pi^2 u_1 u_2} = 4R^2,$$

and so, if ρ is chosen to be sufficiently small and R is sufficiently large, then

$$\frac{\sin(2\pi R u_1)\sin(2\pi R u_2)}{\pi^2 u_1 u_2} > 1/2$$

for every $|\underline{u}| \ll \rho$. Combining this with (13.13) leads us to conclude that

$$\frac{\sin(2\pi R u_1)\sin(2\pi R u_2)}{\pi^2 u_1 u_2} V(u_1, u_2) > \rho^{n-2}/4 > 0 \, \forall \, |\underline{u}| \ll \rho$$

for such a choice of ρ and R. Hence, we have $\mathfrak{J}(R) \gg_\rho 1$ by (13.12) and so we may conclude that $\mathfrak{J} > 0$ by (13.9).

13.3 Proving that $\mathfrak{S} > 0$

In Section 13.1, we proved that

$$\mathfrak{S} := \sum_{q=1}^{\infty} q^{-n} \sideset{}{^*}\sum_{\underline{a}} S_{\underline{a},q} = \prod_{p \text{ prime}} \Big(1 + \sum_{k=1}^{\infty} p^{-kn} S_{\underline{a},p^k}\Big) \tag{13.14}$$

was absolutely convergent provided that $n - \sigma \geq 35$, but in order to complete the proof of Theorem 3.2, we must show that $C_X = \mathfrak{S}\mathfrak{J} > 0$. We proved that $\mathfrak{J} > 0$ in the previous section and so all that is left to do is verify that $\mathfrak{S} > 0$. We will use the assumption that $\sigma := \dim \mathrm{Sing}(F, G) = -1$ throughout this section. With a bit of work, the argument provided by Davenport in the case of one cubic form in [9, Chapter 17] is adaptable to a system of two cubic forms, and so we will work through an analogue of his proof in this section.

Firstly, we note that our proof that $\mathfrak{S}$ converges absolutely gives us

$$|a(q)| := |q^{-n} \sideset{}{^*}\sum_{\underline{a}} S_{\underline{a},q}| \ll q^{-1-\phi} \tag{13.15}$$

for every $q \in \mathbb{N}$ and for some $0 < \phi < 1/16$. In particular, we have

$$\Big| \sum_{k=1}^{\infty} p^{-kn} \sideset{}{^*}\sum_{\underline{a}} S_{\underline{a},p^k} \Big| = \sum_{k=1}^{\infty} |a(p^k)| \ll p^{-1-\phi}\Big(1 + \sum_{k=2}^{\infty} p^{-k-k\phi}\Big) \ll p^{-1-\phi},$$

since the value of the right-most sum is between 0 and 1 for every prime p. Hence, if we let

$$\chi(p) := 1 + \sum_{k=1}^{\infty} p^{-kn} S_{\underline{a},p^k},$$

then

$$|\chi(p) - 1| \ll p^{-1-\phi}$$

for every p. It is easy to check (by taking logs for example) that this implies that

$$1/2 < \prod_{p > p_0} \chi(p) < 3/2 \tag{13.16}$$

for some $p_0 > 1$. Hence, we only need to show that $\chi(p) > 0$ for every prime $p \leq p_0$. In order to do this, we will need the following lemma of Davenport [9, Lemma 5.3]:

Lemma 13.4. *Let*

$$M(q) := \{\underline{x} \bmod q \ : \ F(\underline{x}) \equiv G(\underline{x}) \equiv 0 \bmod q\}.$$

Then, we have

$$\chi(p) = \lim_{k \to \infty} \frac{\# M(p^k)}{p^{k(n-2)}}$$

The statement of the lemma is given in the context of Waring's problem, but the proof directly translates into this context (as noted in [9, Chapter 17]). Therefore, our task of proving that $\chi(p) > 0$ is equivalent to showing that $\# M(p^k) \gg_p p^{k(n-2)}$. Note that the implied constant is allowed to depend on p because $p \leq p_0$, and so this implied constant will be uniformly bounded from below for every such p.

We begin by recalling an assumption about F and G that was made in Chapter 3, namely $F(\underline{x}) \equiv G(\underline{x}) \equiv 0 \bmod p^k$ has non-trivial solutions for every $p, k \in \mathbb{N}$, p prime. We also note that we must have some $l \in \mathbb{N}$ and some $\underline{x} \in (\mathbb{Z}/p^l\mathbb{Z})^n$ such

that

$$F(\underline{x}) \equiv G(\underline{x}) \equiv 0 \bmod p^k, \quad \mathrm{Rank}_{p^l}\begin{pmatrix}\nabla F(\underline{x})\\ \nabla G(\underline{x})\end{pmatrix} = 2. \tag{13.17}$$

Indeed if for every $\underline{x}$ such that $F(\underline{x}) \equiv G(\underline{x}) \equiv 0 \bmod p^k$, we have $\nabla F(\underline{x}) \equiv \underline{0} \bmod p^k$ or $\nabla G(\underline{x}) \equiv \lambda \nabla F(\underline{x}) \bmod p^k$, this is equivalent to there being an $\underline{x} \in \mathbb{Q}_p$ such that $\underline{x} \in \mathrm{Sing}(F, G)$, but $\dim \mathrm{Sing}(F, G) = -1$.

We will now prove that $\#M(p^k) \gg_p p^{k(n-2)}$ by following a Hensel Lemma-type argument, similar to what is used in [9, Chapter 17]. We wish to count the number of solutions $\underline{x} \in M(p^k)$, and to do this we will consider

$$N(p^{k+l}) := \{\underline{x} \in M(p^{k+l}) : \mathrm{Rank}_{p^l}(\nabla F(\underline{x}), \nabla G(\underline{x})) = 2\}, \tag{13.18}$$

where $l \in \mathbb{N}$ is defined as in (13.17). We clearly have that $\#M(p^{k+l}) \geq \#N(p^{k+l})$, and so it suffices to prove that $\#N(p^{k+l}) \gg_p p^{k(n-2)}$. In this case, we aim to verify the following statement by induction:

$$\#N(p^{2l-1+k}) \gg_{p,l} p^{k(n-2)} \; \forall k \in \mathbb{N} \cup \{0\}. \tag{13.19}$$

We note that $\#N(p^{2l-1}) \geq \#N(p^l) \geq 1$ if we define l as in (13.17), and so $k = 0$ is automatically true. Hence

For the induction step, we will assume that (13.19) is true for every $k \leq j$ and consider $j + 1$.

$\underline{x} = \underline{u} + p^{2l+j}\underline{v}$ where $\underline{x} \in \mathbb{Z}/p^{2l+j}\mathbb{Z}$, $\underline{u} \in \mathbb{Z}/p^{j+l}\mathbb{Z}$, and $\underline{v} \in \mathbb{Z}/p^l\mathbb{Z}$, then we must have $\underline{u} \in M(p^{l+j})$. In particular this implies that $F(\underline{u}) = p^{l+j}d_1$, $G(\underline{u}) = p^{l+j}d_2$ for some $d_1, d_2 \in \mathbb{Z}^n$. Hence

$$F(\underline{x}) \equiv G(\underline{x}) \equiv 0 \bmod p^{2l+j} \tag{13.20}$$

$$\iff F(\underline{u}) + p^{l+j}\underline{v} \cdot \nabla F(\underline{u}) \equiv G(\underline{u}) + p^{l+j}\underline{v} \cdot \nabla G(\underline{u}) \equiv 0 \bmod p^{2l+j}$$

$$\iff d_1 + \underline{v} \cdot \nabla F(\underline{u}) \equiv d_2 + \underline{v} \cdot \nabla G(\underline{u}) \equiv 0 \bmod p^l. \tag{13.21}$$

Since $\#N(p^{l+j}) \geq \#N(p^l) \geq 1$ and $N(p^{l+j}) \subset M(p^{l+j})$, we may restrict our consid-

erations to $\underline{u} \in N(p^{l+j}) \neq \phi$, and so $\text{Rank}_l(\nabla F(\underline{x}), \nabla G(\underline{x})) = 2$. Hence, for every such $\underline{u}$, we must have $p^{l(n-2)}$ solutions for $\underline{v}$ in (13.21). By the induction hypothesis, when $j \geq l$, $\#N(p^{l+j}) = \#N(p^{2l-1+(j+1-l)}) \gg_{p,l} p^{(j+1-l)(n-2)}$, and so

$$\#N(p^{2l+j}) \geq \#N(p^{l+j})p^{l(n-2)} \gg_{p,l} p^{(j+1)(n-2)}.$$

Alternatively, when $j < l$

$$\#N(p^{l+j}) \geq \#N(p^l) \geq 1 = p^{-j(n-2)}p^{j(n-2)} \geq p^{-l(n-2)}p^{j(n-2)},$$

and so we still have

$$\#N(p^{2l+j}) \gg_{p,l} p^{(j+1)(n-2)},$$

which completes the induction step. Finally, we note that l depends only on p (since l is chosen to be the smallest value such that (13.17) is true) and so we actually have shown

$$\#N(p^{2l+k}) \gg_p p^{k(n-2)}$$

for every p and every $k \geq 0$. Hence

$$\#M(p^k) \geq \#N(p^k) \gg_p p^{k(n-2)}$$

for every $k \geq 2l$ and so $\chi(p) \gg_p 1$ by Lemma 13.4. In particular, this implies that exists some constant $c(p_0)$ such that $\chi(p) \geq c(p_0) > 0$ for every $p \leq p_0$. Combining this with (13.14) and (13.16) gives

$$\mathfrak{S} = \prod_{p \text{ prime}} \chi(p) = \prod_{p \leq p_0 \text{ prime}} \chi(p) \prod_{p > p_0 \text{ prime}} \chi(p) > \frac{1}{2} c(p_0)^{\hat{p_0}} > 0,$$

where $\hat{p_0}$ is defined to be the number of primes less than or equal to p_0. This completes the proof that $\mathfrak{S} > 0$ when $\sigma = -1$.

Chapter 14

Future Plans

There are several projects that naturally arise out of the work that I have done during the PhD. I also have some ideas that I would like to investigate which have been inspired by the contents of this thesis, but are not directly related. We will briefly discuss each of these in turn.

Optimisation Algorithm

Firstly, I would like to further develop my algorithm which automates the optimisation step of the circle method. The algorithm that I have developed should work for any circle method type problem provided that the degree and the number of forms are fixed, but right now, the algorithm is not in a user-friendly form, and so most people in the area would have difficulty using it. Therefore, my first goal would be to develop a graphical user interface (GUI) for the algorithm to enable researchers who are not experienced with Python to use it.

If I were to do this, then the algorithm will have a significant impact on this area of research: Firstly, it would save a significant amount of research time, as the circle method is a relatively active area of research. If everybody had access to the algorithm, they would be saving several days of research time (almost) every time they write a paper using the circle method, and this should translate to potentially

months of research days saved every year, which will noticeably accelerate progress in the long run. Furthermore, papers will become shorter and simpler, which will make the area more accessible, and will make it easier for academics in the field to publish.

Besides building a GUI, there are several other facets of the algorithm that I would like to improve: Firstly, the current version of the algorithm is extremely space inefficient. In particular, the algorithm required over 20GB of RAM to run for my two cubics optimisation, which is far too much, and so addressing this would be the next priority. I believe that there is a simple change which I can implement to reduce the RAM required to less than 1MB for pretty much any circle method optimisation problem, so improving the space efficiency should not be too difficult.

The current algorithm is also very time inefficient. A few months after building the algorithm, I realised that there is a much faster way to determine a set of points which contain $\mathrm{Crit}(F, D) \cap D$ by representing the data as matrices (instead of list of lists), and performing row reduction. I hope that implementing this change will also simplify the code somewhat. If the algorithm is still relatively slow, then I will also consider rebuilding it on C++ instead of Python.

Finally, I would like to investigate ways to remove the restriction that the degree and number of forms must be fixed, as this would make the algorithm applicable in every circle method type problem. I think this will be the most challenging thing to improve, but it should be possible to achieve this in principle (I will consider collaborating with more experienced programmers for this if there is interest).

Generalisation of Two Cubics

Besides improving the algorithm, the most natural problem which arises from this thesis is trying to apply the techniques developed to a more general problem. For example, instead of considering two cubics, we could consider two forms of degree

$d \geq 3$. In principle, it should be possible to use the results from this thesis as an inductive base case, since we would be able to perform van der Corput differencing $d-2$ times to get down to two quadrics, and at this point, the exponential sum bounds that we have found would be applicable. I anticipate that this will be quite challenging unless the algorithm can also be generalised to cope with forms which do not have fixed degree. This is because the optimisation process for forms with arbitrary degree d would be even more complicated than the case when $d = 3$, and so I suspect it would impractical to perform this process by hand.

An even more ambitious project would be to try working with R forms of degree d, and then find a way to capitalise on the Kloosterman refinement found in this thesis while using Weyl differencing. If one were to find a way to do this which was also compatible with the recent work from Meyerson [24] – [25], then this would lead to a further improvement to Birch's Theorem. This is likely to be very difficult however.

Applications to Three Quadrics

Another natural question which arises from my PhD is whether or not it is possible to improve on Birch's result for three quadrics, as – now that two cubics has been improved – this is the only case from Birch's paper that is not improved upon by Myerson's work [26]. This is a particularly interesting special case due to the fact that it is not viable to apply Poisson summation directly, and if one applies Poisson summation after doing van der Corput differencing once, one cannot expect square root cancellations due to differencing giving linear polynomials. It would be very interesting to see which ideas can be applied effectively in this setting.

Jutila's Big Blocks

At the start of the PhD, I started developing a 2-dimensional analogue to Jutila's "big blocks". This technique was first used in a different context in 1992 [17] before

being applied to the circle method in 1997 [18]. Essentially, Jutila's idea was to take much larger intervals about our rational points when setting up the circle method (relative to the usual Dirichlet intervals) in order to cover the unit square many times, and then divide whatever bound we get by the number of times the unit square is covered. The reason for doing this is that it introduces another way of getting Kloosterman refinement, and so in principle this should further improve our minor arcs bounds for two cubics if we were to combine Jutila's big blocks with the methods used in this thesis.

I have checked that combining this idea with the work I have already done should save at least one additional variable in the case of two cubics, and so I would like to write a follow-up paper on this. In terms of impact, better methods are known for the 1-dimensional circle method (and so it cannot be used), and it is impossible to save many variables using big blocks, but the core idea is very general and can be used in almost any context to potentially save extra variables when using the R-dimensional circle method, for $R > 1$.

"Multiple" Averaged van der Corput Differencing

I would also like to investigate a potential improvement to the Averaged van der Corput Differencing method in the case where one differences at least twice. In particular, my preliminary investigations lead me to believe that it may be possible to perform averaging over the integral after every differencing step instead of just the first step. If one starts with R forms of the same degree, then Averaged van der Corput Differencing gives a saving of $(P/H)^{R/2}$, whilst my proposed Multiple Averaged van der Corput Differencing would presumably give a saving of $(P/H)^{R(2^D-1)/2^D}$, where D is the number of differencing steps. One quintic form or two quartic forms would be logical candidates to apply this idea to, since we must difference twice before we can apply Poisson summation in these settings. This investigation is still in its early stages, but this would be a very interesting, and very general result if

averaging over the integral multiple times is possible.

Function Fields

Applying the methods used in this thesis to the function fields setting is another potential avenue of exploration, as every idea that I have discussed in the thesis in relation to the circle method should have an analogue in this context. There is a version of Kloosterman refinement in the function fields setting, so to get optimal results, one would need to combine my work with the work done by Vishe in [28].

Going in the reverse direction, the version of Kloosterman refinement developed by Vishe in the function fields setting currently does not have an analogue for complete intersections over $\mathbb{Q}$, and so it would be interesting to see whether it is possible to find this.

Appendix A: Mathematica Code

Here, we will include the Mathematica code that verifies our minor arcs bound. An executable version of this can be found at [27].

In[1]:=
```
ϵ = 1 / 10 000
Δ = 1 / 7 - 1 / 1000

AVDCPoissonBound[ϕ_, τ_, ϕ3_, ϕ4_, ϵ_, n_] :=
 n - 1 + 5 ϕ / 2 + (2 - n) / 2 Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ] + 2 Max[-2 - Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ], τ] +
  1 / 2 Max[ϕ, (1 - n) ϕ / 2 + n Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ] +
      (n - 1) Max[0, -1 + ϕ, ϕ + (τ + Max[10 / (n-2) + ϵ, (2+6 ϕ) / (n+2) + ϵ]) / 2],
     (1 - n) ϕ / 2 + (n / 3 - 1 / 2) ϕ3 + (1 - n) ϕ4 / 2 + n Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ]]]

PVDCPoissonBound[ϕ_, τ_, ϕ3_, ϕ4_, ϵ_, n_] :=
 n + 5 ϕ / 2 + 2 τ - n / 2 Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ] + 1 / 2 Max[ϕ, (1 - n) ϕ / 2 +
      n Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ] + (n - 1) Max[0, -1 + ϕ, ϕ + (τ + Max[10 / (n-2) + ϵ, (2+6 ϕ) / (n+2) + ϵ]) / 2],
     (1 - n) ϕ / 2 + (n / 3 - 1 / 2) ϕ3 + (1 - n) ϕ4 / 2 + n Max[10 / (n - 2) + ϵ, (2 + 6 ϕ) / (n + 2) + ϵ]]

AVDCWeylBound[ϕ_, τ_, ϕ3_, ϕ4_, ϵ_, n_] :=
 n - 1 + 3 ϕ - 2 ϕ3 / 3 - 3 ϕ4 / 4 + (3 - n) / 2 Max[ϕ / 6, (2 + ϕ + τ) / 5] + 2 Max[-2 - Max[ϕ / 6, (2 + ϕ + τ) / 5]]
PVDCWeylBound[ϕ_, τ_, ϕ3_, ϕ4_, ϵ_, n_] :=
 n + 3 ϕ + 2 τ - 2 ϕ3 / 3 - 3 ϕ4 / 4 + (1 - n) / 2 Max[ϕ / 6, (2 + ϕ + τ) / 5]

WeylWeylBound[ϕ_, τ_, ϕ3_, ϕ4_, ϵ_, n_] :=
 n + 3 ϕ + 2 τ - 2 ϕ3 / 3 - 3 ϕ4 / 4 + (n - 1) / 16 Max[2 ϕ + 2 τ, -ϕ + Min[0, -3 - τ]]

MinorArcsBound[ϕ_, τ_, ϕ3_, ϕ4_, ϵ_, n_] := Min[AVDCPoissonBound[ϕ, τ, ϕ3, ϕ4, ϵ, n],
  PVDCPoissonBound[ϕ, τ, ϕ3, ϕ4, ϵ, n], AVDCWeylBound[ϕ, τ, ϕ3, ϕ4, ϵ, n],
  PVDCWeylBound[ϕ, τ, ϕ3, ϕ4, ϵ, n], WeylWeylBound[ϕ, τ, ϕ3, ϕ4, ϵ, n]]
```

Out[1]= $\frac{1}{10\,000}$

Out[2]= $\frac{993}{7000}$

In[9]:=
```
For[n = 39, n ≤ 42, n++,
 Print[n, N[Maximize[{MinorArcsBound[ϕ, τ, ϕ3, ϕ4, ϵ, n], Δ ≤ ϕ ≤ 3/2,
      -10 ≤ τ ≤ -3/4 - ϕ, 0 ≤ ϕ3 ≤ ϕ, 0 ≤ ϕ4 ≤ ϕ - ϕ3}, {ϕ, τ, ϕ3, ϕ4}]]]]
```

```
39{32.9982, {ϕ → 1.5, τ → -2.25, ϕ3 → 0., ϕ4 → 0.}}
40{33.9981, {ϕ → 1.5, τ → -2.25, ϕ3 → 0., ϕ4 → 0.}}
41{34.9981, {ϕ → 1.5, τ → -2.25, ϕ3 → 0., ϕ4 → 0.}}
42{35.998, {ϕ → 1.5, τ → -2.25, ϕ3 → 0., ϕ4 → 0.}}
```

In[10]:=
```
For[n = 43, n ≤ 48, n++,
 Print[n, N[Maximize[{MinorArcsBound[ϕ, τ, ϕ3, ϕ4, ϵ, n], 4 (n + 3)/(3 (n - 2)) ≤ ϕ ≤ 3/2,
      -10 ≤ τ ≤ -3/4 - ϕ, 0 ≤ ϕ3 ≤ ϕ, 0 ≤ ϕ4 ≤ ϕ - ϕ3}, {ϕ, τ, ϕ3, ϕ4}]]]]
```

```
43{36.9978, {ϕ → 1.5, τ → -2.5, ϕ3 → 0., ϕ4 → 0.003}}
44{37.881, {ϕ → 1.49206, τ → -2.49206, ϕ3 → 0., ϕ4 → 0.}}
45{38.7597, {ϕ → 1.48837, τ → -2.48837, ϕ3 → 0., ϕ4 → 0.}}
46{39.6389, {ϕ → 1.48485, τ → -2.48485, ϕ3 → 0., ϕ4 → 0.}}
47{40.5185, {ϕ → 1.48148, τ → -2.48148, ϕ3 → 0., ϕ4 → 0.}}
48{41.3986, {ϕ → 1.47826, τ → -2.47826, ϕ3 → 0., ϕ4 → 0.}}
```

In[11]:=
```
For[n = 43, n ≤ 48, n++,
 Print[n, N[Maximize[{MinorArcsBound[ϕ, τ, ϕ3, ϕ4, ϵ, n], Δ ≤ ϕ ≤ 4 (n + 3)/(3 (n - 2)),
      -10 ≤ τ ≤ -3/4 - ϕ, 0 ≤ ϕ3 ≤ ϕ, 0 ≤ ϕ4 ≤ ϕ - ϕ3}, {ϕ, τ, ϕ3, ϕ4}]]]]
```

```
43{36.9978, {ϕ → 1.49593, τ → -2.24876, ϕ3 → 0., ϕ4 → 0.}}
44{37.9183, {ϕ → 0.141857, τ → -2.95271, ϕ3 → 0.140332, ϕ4 → 0.}}
45{38.9064, {ϕ → 0.141857, τ → -2.95271, ϕ3 → 0.140364, ϕ4 → 0.}}
46{39.8946, {ϕ → 0.141857, τ → -2.95271, ϕ3 → 0.140395, ϕ4 → 0.}}
47{40.8827, {ϕ → 0.141857, τ → -2.95271, ϕ3 → 0.140424, ϕ4 → 0.}}
48{41.8709, {ϕ → 0.141857, τ → -2.95271, ϕ3 → 0.140453, ϕ4 → 0.}}
```

In[12]:=
```
For[n = 39, n ≤ 48, n++,
 Print[n, N[Maximize[{MinorArcsBound[ϕ, τ, ϕ3, ϕ4, ϵ, n], 0 ≤ ϕ ≤ Δ,
      -3 + Δ ≤ τ ≤ -3/4 - ϕ, 0 ≤ ϕ3 ≤ ϕ, 0 ≤ ϕ4 ≤ ϕ - ϕ3}, {ϕ, τ, ϕ3, ϕ4}]]]]
```

39{32.9744, {$\phi \to 0.044185$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

40{33.9628, {$\phi \to 0.0442594$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

41{34.9512, {$\phi \to 0.0443304$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

42{35.9396, {$\phi \to 0.044398$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

43{36.928, {$\phi \to 0.0444627$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

44{37.9164, {$\phi \to 0.0445245$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

45{38.9048, {$\phi \to 0.0445837$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

46{39.8931, {$\phi \to 0.0446404$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

47{40.8815, {$\phi \to 0.0446947$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

48{41.8698, {$\phi \to 0.0447469$, $\tau \to -2.85814$, $\phi 3 \to 0.$, $\phi 4 \to 0.$}}

Bibliography

[1] B. J. Birch. Forms in many variables. Proc. Roy. Soc. Ser. A, 265:245–263, 1961/1962.

[2] T. D. Browning, R. Dietmann, and D. R. Heath-Brown. Rational points on intersections of cubic and quadric hypersurfaces. J. Inst. Math. Jussieu, 14(4):703–749, 2015.

[3] T. D. Browning and D. R. Heath-Brown. Rational points on quartic hypersurfaces. J. reine angew. Math., 629:37–88, 2009.

[4] T. D. Browning and D. R. Heath-Brown. Forms in many variables and differing degrees. J. Eur. Math. Soc., 19(2):357–394, 2017.

[5] T. D. Browning and S. M. Prendiville. Improvements in Birch's theorem on forms in many variables. J. reine angew. Math., 2017(731):203–234, 2017.

[6] J. Brüdern and T. D. Wooley. Cubic moments of fourier coefficients and pairs of diagonal quartic forms. Journal of the European Mathematical Society, 17(11):2887–2901, 2015.

[7] J. Brüdern and T. D. Wooley. The Hasse principle for systems of diagonal cubic forms. Mathematische Annalen, 364(3-4):1255–1274, 2016.

[8] J. Brüdern and T. D. Wooley. Arithmetic harmonic analysis for smooth quartic weyl sums: three additive equations: three additive equations. Journal of the European Mathematical Society, 20(10):2333–2356, 2018.

[9] H. Davenport. Analytic methods for Diophantine equations and Diophantine inequalities. Cambridge Mathematical Library. Cambridge University Press, Cambridge, second edition, 2005. With a foreword by R. C. Vaughan, D. R. Heath-Brown and D. E. Freeman, Edited and prepared for publication by T. D. Browning.

[10] M. A. Hanselmann. Rational points on quartic hypersurfaces. PhD thesis, Ludwig-Maximilians-Universität München, 2012.

[11] D. R. Heath-Brown. Cubic forms in ten variables. Proc. London Math. Soc. (3), 47(2):225–257, 1983.

[12] D. R. Heath-Brown. A new form of the circle method, and its application to quadratic forms. J. reine angew. Math., 481:149–206, 1996.

[13] D. R. Heath-Brown. Cubic forms in 14 variables. Invent. Math., 170(1):199–230, 2007.

[14] D. R. Heath-Brown and L. B. Pierce. Simultaneous integer values of pairs of quadratic forms. J. reine angew. Math., 727:85–143, 2017.

[15] C. Hooley. On the representations of a number as the sum of four cubes. I. Proc. London Math. Soc. (3), 36(1):117–140, 1978.

[16] Christopher Hooley. On nonary cubic forms. J. reine angew. Math., 386:32–98, 1988.

[17] M. Jutila. Transformations of exponential sums. In Proc. Amalfi Conf. on Analytic Number Theory, pages 263–270, 1992.

[18] M. Jutila. Sieve Methods, Exponential Sums, and their Applications in Number Theory: A variant of the circle method, pages 245–254. London Mathematical Society Lecture Note Series. Cambridge University Press, 1997. Edited by G. R. H. Greaves, G. Harman, and M. N. Huxley.

[19] H. D. Kloosterman. On the representation of numbers in the form $ax^2 + by^2 + cz^2 + dt^2$. Acta Mathematica, 49(3):407–464, 1927.

[20] S. A. Lee. Birch's theorem in function fields. Preprint, 2011. arXiv:1109.4953.

[21] O. Marmon and P. Vishe. On the Hasse principle for quartic forms. Duke Math J, 168(14):2727–2799, 2019.

[22] Oscar Marmon. The density of integral points on complete intersections. Q. J. Math., 59(1):29–53, 2008. With an appendix by Per Salberger.

[23] R. Munshi. Pairs of quadrics in 11 variables. Compos. Math., 151(7):1189–1214, 2015.

[24] S. L. R. Myerson. Systems of forms in many variables. arXiv:1709.08917.

[25] S. L. R. Myerson. Systems of cubic forms in many variables. J. reine angew. Math., 2019(757):309–328, 2017.

[26] S. L. R. Myerson. Quadratic forms and systems of forms in many variables. Invent. Math., 213:205–235, 2018.

[27] M. J. Northey. The final optimisation algorithm along with mathematica code. `https://github.com/MJNorthey/Circle-Method-Algorithm/releases/tag/v1.0`.

[28] P. Vishe. Rational points on complete intersections on $\mathbb{F}_q(t)$. arXiv:1907.07097.