# Topics in Combinatorics and Additive Number Theory




James Aaronson
Magdalen College
University of Oxford

A thesis submitted for the degree of
*Doctor of Philosophy*
Trinity 2020

# Acknowledgements

Firstly, I would like to thank my advisor, Ben Green, for his continued support and guidance, and for providing a healthy supply of interesting problems to think about.

I would also like to thank Carla Groenland and Tom Johnston for being excellent collaborators.

I am grateful to the anonymous referees of the paper [1] for their helpful comments; that paper forms part of Chapter 3 of this thesis.

I would also like to thank Jakub Konieczny, Aled Walker, Sofia Lindqvist, Dan Altman and Zachary Chase for many valuable conversations.

My work was funded by a generous grant from the EPSRC; I am very grateful for their support.

Last, but not least, I would like to thank my family for their endless encouragement.

# Abstract

In this thesis, we will consider five problems with an additive combinatorial flavour.

In Chapter 2, we will consider bounds for the arithmetic removal lemma in abelian groups, and show that good bounds would hold as a consequence of good bounds on a related problem.

In Chapter 3, we will discuss sets of integers with many solutions to a linear equation, and prove various results about the maximal number of solutions for various equations.

In Chapter 4, we will consider special subspaces of a finite vector space called Cyclically Covering Subspaces, and discuss bounds for the smallest such subspaces. The work in this chapter answers a question asked by Cameron, and is joint with Carla Groenland and Tom Johnston.

In Chapter 5, we will prove that, among functions supported on a Hamming Sphere with a fixed $\ell^2$ norm, the one which maximises the additive energy is constant, proving a conjecture of Kirshner and Samorodnitsky.

In Chapter 6, we will present a possible strengthening of the celebrated Polynomial Freiman-Ruzsa conjecture, and show why such a strengthening cannot hold.

# Contents


**1 Introduction** **1**

**2 A connection between matchings and removal in abelian groups** **12**

2.1 Introduction . . . . . 12

2.2 Theorem 2.1.1 for $N$ prime . . . . . 14

2.3 Theorem 2.1.1 for general $N$ . . . . . 17

**3 Maximising the number of solutions to a linear equation in a set of integers** **19**

3.1 Introduction . . . . . 19

3.2 The elementary approach to Theorem 3.1.4 . . . . . 24

3.2.1 Additive combinatorial lemmas . . . . . 24

3.2.2 Proof of Theorem 3.1.4 . . . . . 27

3.3 Proving Lemma 3.1.5 via arithmetic regularity . . . . . 32

3.3.1 Finding a structured subset . . . . . 33

3.3.2 Applying the arithmetic regularity lemma . . . . . 36

3.3.3 Simplifying the local problem . . . . . 42

3.4 Completing the proof of Theorem 3.1.4 . . . . . 47

3.5 Equations in more than three variables . . . . . 50

3.6 Generalising Lemma 3.1.5 to more than three variables . . . . . 58

3.7 Systems of more than one equation . . . . . 61

3.8 The case $x + 2y + z = 0$ . . . . . 63

3.9 Equations of the form $x + y = kz$ . . . . . 65

3.9.1 The set of numbers of the form $\gamma_{a_1,a_2,a_3}$ . . . . . 68

**4 Cyclically covering subspaces in $\mathbb{F}_2^n$** **70**

4.1 Introduction . . . . . 70

4.2 Multiplicative bounds . . . . . 73

4.3 Primitive roots and a structural conjecture . . . . . 77

4.4 Partial progress towards conjecture . . . . . 83
4.4.1 A further reformulation of the problem . . . . . 83
4.4.2 Restricting the edges of the digraph . . . . . 85
4.4.3 Eliminating large girth . . . . . 88
4.5 Proof of the main result . . . . . 95
4.6 Generalisation to arbitrary finite fields . . . . . 104

**5 Functions of large additive energy supported on a Hamming Sphere 110**
5.1 Introduction . . . . . 110
5.2 Proof of Theorem 5.1.1 . . . . . 111

**6 A counterexample to a strong variant of the Polynomial Freiman-Ruzsa conjecture 115**
6.1 Introduction . . . . . 115
6.2 The construction . . . . . 116

**Bibliography 118**

# Chapter 1

# Introduction

In this thesis, we will consider several problems with a combinatorial flavour, which will be discussed in the next five self-contained chapters. In this introductory chapter, we will give an overview of the problems discussed.

The first of these problems concerns the question of removal in abelian groups. This question has an obvious analogue in the graph setting: we are given a graph $G$ on $N$ vertices, and wish to delete some edges in order to remove all of the triangles. Ruzsa and Szemerédi [50], using Szemerédi's regularity lemma, proved the following.

**Theorem** (Triangle removal in graphs)**.** *For any $\varepsilon > 0$, there is a $\delta > 0$ with the property that any graph with no more than $\delta N^3$ triangles can be made triangle-free by removing no more than $\varepsilon N^2$ of the edges.*

The natural question in this setting is that of how strong we can make the bounds on $\delta$. Ruzsa and Szemerédi's original proof only gave tower-type bounds (more specifically, $\delta^{-1}$ is given as a tower of height polynomial in $\varepsilon^{-1}$) coming from the use of Szemerédi's regularity lemma. For more information on this problem, see [13].

The problem we consider is concerned with the analogous problem in the context of abelian groups. In this situation, we are given an abelian group $G$ and a set $A \subseteq G$, and define a *triangle* to be a solution to the equation $x + y + z = 0$.

Our goal now is to delete elements of $A$ in order to remove all triangles from $G$. Green's arithmetic triangle removal lemma, proven using the arithmetic regularity lemma introduced in [25], says the following.

**Theorem** (Triangle removal in abelian groups)**.** *Let $G$ be an abelian group. For any $\varepsilon > 0$, there is a $\delta > 0$ with the property that any subset with no more than $\delta N^3$ triangles can be made triangle-free by removing no more than $\varepsilon N^2$ of the elements.*

Note that this theorem is strongly related to the graph version via the concept of a *Cayley graph*. Recall that a Cayley graph on a group $G$ generated by $A$ is simply the graph with vertex set $G$, and where $a$ and $b$ are joined by an edge if $a-b$ or $b-a$ is an element of the generating set $A$. A directed Cayley graph has us join an edge directed from $a$ to $b$ if and only if $b-a$ is an element of $A$. Thus, a triangle in a set $A$ corresponds directly to a directed triangle in the Cayley graph generated by $A$. Indeed, Král, Serra and Vena use this formulation in [42], combined with a directed variant of the graph removal lemma due to Alon and Shapira [2, Lemma 4.1], to prove the triangle removal lemma in groups.

Green's proof of the triangle removal lemma uses the arithmetic regularity lemma, and thus required tower-type bounds, analogous with the graph version. Indeed, the bounds achieved by the proof of Král, Serra and Vena are also of tower type, due to their reliance on the graph removal lemma. Our goal is to improve the bounds on $\delta$ in the group triangle removal lemma.

In the special case $G = \mathbb{F}_p^n$ for a fixed prime $p$, Fox and Lovász [20] proved a much stronger bound, namely that $\delta^{-1}$ is bounded by a polynomial in $\varepsilon^{-1}$. Their argument followed on from work by Croot-Lev-Pach [15] and subsequent work by Ellenberg-Gijswijt [18] on the cap set problem.

**Theorem** (Ellenberg-Gijswijt)**.** *If $A \subseteq \mathbb{F}_p^n$ is a set with no three-term arithmetic progression, then $|A| \leqslant p^{cn}$ for some constant $c < 1$.*

Define an *additive matching* to be a collection of triples $(x_i, y_i, z_i)$ such that $x_i + y_j + z_k = 0$ if and only if $i = j = k$. These are also called tricolored sum-free sets, and can be represented by $(X, Y, Z)$, where $X = \{x_i\}$, $Y = \{y_i\}$ and $Z = \{z_i\}$. Kleinberg [41] gave a polynomial upper bound for the size of an additive matching in $G$ in the case that $G = \mathbb{F}_2^n$ using arguments related to those of Croot-Lev-Pach and Ellenberg-Gijswijt, and Blasiak et al. [6] extended this to $\mathbb{F}_q^n$ for a fixed prime power $q$. The argument by Fox and Lovász made use of these results to prove the polynomial bounds on removal.

In Chapter 2, we will consider the same question, but in cyclic groups. We will not be able to obtain an unconditional result without better bounds on the size of an additive matching in a cyclic group. However, we will be able to adapt the argument of Fox and Lovász to show that, if we had good bounds on the size of an additive matching, then we could get good bounds on removal. For instance, we have the following.

**Theorem.** *Suppose we could prove that any additive matching in $\mathbb{Z}/N\mathbb{Z}$ has density at most $\exp(-c\sqrt{\log N})$ for some constant $c$.*

*Then, the group triangle removal lemma holds with*

$$\varepsilon \ll \exp(-c_1\sqrt{\log 1/\delta})$$

*for some constant $c_1$.*

Note that the proposed bound $\exp(-c\sqrt{\log N})$ is not unreasonable; indeed, it is the bound obtained from Behrend's construction [4] of sets with no three terms in arithmetic progression, which is plausibly the upper bound for the size of an additive matching.

The arguments we use to prove the above theorem are adapted from those used by Fox and Lovász. The idea behind their proof strategy is as follows. Given a subset $S \subseteq \mathbb{F}_p^n$ for which triangle removal requires deletion of too many points, we can find a large subset of $S$ in which no point is contained in too many triangles. Then, we pass to a random subspace of $\mathbb{F}_p^n$, and show that with positive probability that subspace contains a large additive matching.

The natural analogue of this idea in the case of $\mathbb{Z}/N\mathbb{Z}$ would be passing to a random progression. Unfortunately, this does not work, essentially because $\mathbb{Z}/N\mathbb{Z}$ does not have 'enough' progressions. To get around this problem, we instead pass to a triple of progressions, and consider triangles with one element in each of the progressions.

Perhaps in the other direction, suppose we want to find a set $S$ of integers of size $N$ with as many solutions as possible to the equation $x + y + z = 0$. The answer to this is fairly well-known; it is best to take $S$ to be the set of integers in the interval $[1 - \lceil N/2 \rceil, \lfloor N/2 \rfloor]$, for a total of $\frac{3}{4}N^2 + O(N)$ solutions. This was proven by Hardy and Littlewood [32] and Gabriel [21] in the case of odd $N$, and extended to even $N$ by Lev in [43].

The ideas behind their approaches involve rearrangement inequalities, which involve maximising a sum over the terms of a sequence or sequences over all rearrangements of that sequence (or sequences). A prototypical example of the kind of inequalities we are interested in is as follows.

**Theorem** (Rearrangement inequality)**.** *Suppose we have sequences $(a_i)_{i=1}^n$ and $(b_i)_{i=1}^n$ of nonnegative reals. Then, among all reorderings of the $a_i$ and $b_i$, the value of the sum $\sum_i a_i b_i$ is maximised when both sequences are ordered so as to be increasing.*

It is fairly natural that inequalities of this type should be useful for this type of problem. For instance, given a set $A$ of integers, we may associate to it a sequence $(a_i)_{i\in\mathbb{Z}}$ by $a_i = 1$ if $i \in A$ and $a_i = 0$ otherwise. Rearranging the order of the terms of the sequence simply corresponds to choosing a different set $A$ of the same size, and the number of solutions to $x + y + z = 0$ in $A$ can be written as $\sum_{i+j+k=0} a_i a_j a_k$, and so maximising the number of solutions in $A$ to $x+y+z = 0$ is the same as maximising the given sum over all rearrangements of the sequence $(a_i)$. Rearrangement inequalities are discussed in more detail in [33, Chapter 10], which inspires some of the arguments we will use.

We may wish to ask the same question about other equations. One natural example to take would be the equation $x - 2y + z = 0$; in other words, the equation for a three-term arithmetic progression. It was shown by Green and Sisask in [27, Theorem 1.2] that the maximal number of solutions for a set of size $N$ is again achieved by an interval of length $N$, for a total of $\frac{1}{2}N^2 + O(N)$ solutions.

*Remark.* We say the equation $x - 2y + z = 0$ is *translation-invariant*. This means that its solution set consists of the translates and scalings of an affine pattern (in this case $\{0, 1, 2\}$). Equivalently, the sum of the coefficients of the variables is 0.

By analogy with this, Bhattacharya, Ganguly, Shao and Zhao considered longer arithmetic progressions; in [5, Theorem 2.4], they proved that the number of $k$ term arithmetic progressions in a set $S$ of $n$ integers is maximised when $S$ is an interval.

Motivated by this example, Ganguly asked [22] about other affine patterns; in particular, finding sets $S$ with as many affine copies of $\{0, 1, 3\}$, or solutions to $x+2y = 3z$, as possible. In this case, such a result would certainly be less clean; for instance, there are more solutions to $x + 2y = 3z$ in the set $\{0, 1, 3\}$ than in an interval of length 3.

We can generalise this to consideration of all three-variable linear equations. Given coprime coefficients $a, b$ and $c$, we are interested in the problem of finding sets with as many solutions to $ax + by + cz = 0$ as possible.

The trivial upper bound on the number of solutions is $|S|^2$. This is because, for any choice of $x$ and $y$, there is at most one choice of $z$ such that $ax + by + cz = 0$, namely $z = \frac{-ax-by}{c}$. We are interested in making $T(S)$ as large as possible, among sets of size $|S| = N$.

As discussed, the optimal choice of $S$ is known for the equation $x + y + z = 0$ by the work of Hardy and Littlewood, Gabriel and Lev. The same argument also deals with the equation $x + y = z$ (for which the optimal set is again an interval centered around the origin). The work of Green and Sisask shows that the same set is optimal

for the equation $x - 2y + z = 0$, and Lev and Pinchasi [44] show that the same set is optimal for the equation $x + 2y = z$.

For other equations, much less is known. Indeed, the answer to Ganguly's question about the equation $x+2y = 3z$ is not known. However, we will show that for equations with large coefficients, the interval centered at the origin is far from optimal. Indeed, for any equation $ax + by + cz = 0$ with coprime coefficients, we can find large sets $S$ with at least $\frac{1}{12}|S|^2$ solutions to the equation in $S$. By contrast, the interval centered at the origin of length $|S|$ has no more than $\frac{1}{|c|}|S|^2 + O(|S|)$ solutions, if we assume without loss of generality that $|c| \geqslant |a|, |b|$.

The asymptotic for the maximal number of solutions is therefore between $\frac{1}{12}|S|^2$ and $|S|^2$ for any equation. This raises the natural question of whether these constants $\frac{1}{12}$ and 1 are optimal.

Investigating the upper bound of 1 proves to be fairly trivial. The work of Hardy and Littlewood and Gabriel in fact proves that for any equation, any set $S$ can have at most $\frac{3}{4}|S|^2 + O(|S|)$ solutions, and this is optimal in view of the fact that the equation $x + y + z = 0$ achieves this bound.

However, the lower bound is more interesting. The main result we prove in Chapter 3 is the fact that the lower bound $\frac{1}{12}|S|^2$ cannot be improved. More precisely, we prove the following.

**Theorem.** *For any $\varepsilon > 0$, we can find an equation $ax + by + cz = 0$ with the property that any large set $S$ can have no more than $\left(\frac{1}{12} + \varepsilon\right)|S|^2 + O(|S|)$ solutions.*

In Chapter 3, we will provide two fairly distinct proofs of the above statement. The first argument is elementary, and rests on a structure theorem of Green and Sisask which they proved in [27]. This allows us to partition a candidate set $S$ into highly structured pieces. Once we have done this, we use some elementary arguments to locate where most of the solutions must be with respect to the partition, allowing us to prove the theorem.

Our second approach relies on the arithmetic regularity lemma [25]. The arguments here are more involved than in the elementary proof, but this approach has the advantage of proving a weak structure theorem, of sorts, of optimal sets. More precisely, we describe a restricted class of sets, which must include sets that achieve the correct maximal number of solutions up to an $O(|S|)$ term.

It is worth noting that the two approaches have various other minor advantages and disadvantages. For instance, any proof of our theorem must supply two sets of bounds. The first is a bound on how large the coefficients $a, b$ and $c$ must be,

given $\varepsilon$, to ensure that any large set $S$ can have no more than $(\frac{1}{12}+\varepsilon)|S|^2$ solutions. The second is a bound on how large $|S|$ must be, given $\varepsilon, a, b$ and $c$, such that the same holds. The elementary approach to our theorem provides exponential bounds for both of these quantities, whereas the argument using the regularity lemma gives a polynomial bound on the size of the coefficients, but a tower-type bound on $|S|$. This can be seen as an advantage of the second approach, because the bounds on the coefficients are much more important.

An advantage of the elementary approach is that it can supply a wider class of equations. Indeed, the equations provided by the second approach require all ratios between the coefficients to be large, which means that it cannot provide translation-invariant equations. The first approach only requires the ratios between the coefficients to have large *height*; here, we define the height of a rational number to be the sum of the absolute values of the numerator and denominator. Equivalently, the ratios between the coefficients must have large numerator or denominator. In particular, this allows the first approach to provide translation-invariant equations.

A natural question at this point concerns what happens when we consider linear equations in more variables. Naturally, if $a_1x_1+\cdots+a_kx_k=0$ is a $k$-variable equation, the maximum possible number of solutions in a set $S$ will be at most $|S|^{k-1}$, since each $k-1$-tuple $(x_1,\dots,x_{k-1})$ will give rise to at most one solution. It seems natural to guess that a generalisation of the example giving $\frac{1}{12}|S|^2$ solutions exists, and that is indeed the case. Indeed, if we set

$$\sigma_k=\int_{-\infty}^{\infty}\left(\frac{\sin\pi x}{\pi x}\right)^k dx,$$

then there are large sets $S$ with at least $\frac{\sigma_k}{k^{k-1}}|S|^{k-1}$ solutions. For instance, $\sigma_3=3/4$, which gives rise to the factor of $1/12$ seen earlier. Moreover, we can prove that this constant cannot be improved, using generalisations of the arguments used for the three variable case.

Our third problem concerns the concept of *cyclically covering subspaces*. Given $V\leqslant\mathbb{F}_2^n$ a vector subspace, we say that $V$ is *cyclically covering* if and only if for any vector $v\in\mathbb{F}_2^n$, the coordinates of $v$ can be cyclically shifted to give an element of $V$.

We are interested in how small a cyclically covering subspace can be. We will define $h_2(n)$ to be the codimension of the smallest cyclically covering subspace in $\mathbb{F}_2^n$, and so we are interested in determining, or bounding, $h_2(n)$.

The trivial bounds on $h_2(n)$ are $0\leqslant h_2(n)\leqslant\lfloor\log_2 n\rfloor$, the lower bound coming from the fact that $\mathbb{F}_2^n$ itself is cyclically covering, and the upper bound coming from

an easy counting argument. These are hard to improve. For the upper bound, it has been shown by Cameron, Ellis and Raynaud [11] that if $n$ is of the form $2^k - 1$ then $h_2(n) = \lfloor \log_2 n \rfloor$. For the lower bound, there are examples to show that $h_2(n) \geqslant 2$ when $n \geqslant 5$ is odd (see, for example, [11]).

It is natural to ask whether the lower bound can be improved for large $n$; indeed, Cameron asked [10] whether $h_2(n) \to \infty$ as $n \to \infty$ over the odd numbers.

Our main result, which is joint work with Groenland and Johnston, is conditionally that the answer to Cameron's question is no. Indeed, we prove that $h_2(n) = 2$ for infinitely many values of $n$. More precisely, we show that $h_2(p) = 2$ if $p$ is an Artin prime, which is a prime with 2 as a primitive root. This makes our result conditional on Artin's widely believed conjecture that there are infinitely many such primes. Indeed, strong evidence for Artin's conjecture has been given by work of Hooley [35], which proves Artin's conjecture subject to the Generalised Riemann Hypothesis, and Heath-Brown [34], which proves that there are at most two primes which fail to be a primitive root infinitely often.

The argument we use to prove our result is fairly involved. We will begin by viewing vectors as polynomials in $\mathbb{F}_2[X]/(X^n - 1)$. This has the advantage that the cyclic shift operation is just multiplication by $X$. This allows us to utilise results from Galois theory; in particular, we can use the fact that $n$ being a prime with primitive root 2 implies that the polynomial $1 + X + \cdots + X^{n-1}$ is irreducible.

Once we have done this, we give a second alternate formulation of the problem, which is to equate a vector $v$ with a related directed Cayley graph on the cyclic group $\mathbb{Z}/n\mathbb{Z}$. This allows us to use additive combinatorics results, particularly the Cauchy-Davenport theorem, to prove our result. We will do this in Chapter 4.

The generalisation of cyclically covering subspaces to be defined over $\mathbb{F}_p$ rather than just $\mathbb{F}_2$ was introduced by Cameron, Ellis and Raynaud in [11]. They raised the question of how $h_p(n)$, the codimension of the smallest cyclically covering subspace in $\mathbb{F}_p^n$, behaves as $n$ grows. In fact, if $n$ is a prime with $p$ as a primitive root for an odd prime $p$, we show that $h_p(n) = 0$. This contrasts with the case $p = 2$, where we show that in such cases $h_2(n) = 2$.

Cyclically covering subspaces were investigated by Cameron, Frankl and Cantor in [12], in connection with their work on *homogeneous games*, which were introduced by Isbell in [36] and [37]. The problem involves the following situation. We are given $n$ voters deciding between two candidates, and wish to define a collection of winning sets with the following properties:

- If $A$ and $B$ form a partition of the voters, then exactly one of the two is winning,

- If $A \subseteq B$ and $A$ is winning, then so is $B$,
- For any two voters $v_1$ and $v_2$, there is a permutation of the voters fixing the collection of winning sets which takes $v_1$ to $v_2$.

These are natural conditions one might desire for an election; the first says that exactly one candidate wins, the second says that voters switching their preference to a candidate cannot hurt that candidate and the last condition ensures that any two voters are treated equally.

The central question is to determine whether or not a homogeneous game exists for any given value of $n$. For instance, it is clear that a homogeneous game exists when $n$ is odd, because we can take the winning sets to be exactly the sets of size greater than $n/2$. Isbell conjectured that for each odd $b$, there is an $m(b)$ such that if $a \geqslant m(b)$, then a homogeneous game does not exist for $n = 2^a b$.

The connection between homogeneous games and cyclically covering subspaces is given in [12], where it is shown that $m(b) > h_2(b)$ for any odd integer $b$.

Our fourth problem concerns functions defined on $\mathbb{F}_2^n$. Recall that the *Hamming sphere* of radius $k$ consists of all vectors in $\mathbb{F}_2^n$ with exactly $k$ ones.

Also, recall the Fourier transform on $\mathbb{F}_2^n$. For $\alpha \in \mathbb{F}_2^n$, we define the Fourier-Walsh character $W_\alpha$ by $w_\alpha = (-1)^{\alpha \cdot y}$, where $\alpha \cdot y = \sum \alpha_i y_i$ is the standard dot product. For a function $f : \mathbb{F}_2^n \to \mathbb{R}$, we define the Fourier transform $\hat{f} : \mathbb{F}_2^n \to \mathbb{R}$ as follows:

$$\hat{f}(\alpha) = \mathbb{E}_y f(y) W_\alpha(y).$$

We have the Fourier inversion theorem:

$$f(y) = \sum_\alpha \hat{f}(\alpha) W_\alpha(y),$$

We can define the following norms for functions.

$$||f||_p = (\mathbb{E}_y |f(y)|^p)^{1/p}$$
$$||f||_{U_2}^4 = \left( \mathbb{E}_{\substack{a_1,a_2,a_3,a_4 \in \mathbb{F}_2^n \\ a_1+a_2=a_3+a_4}} f(a_1) f(a_2) f(a_3) f(a_4) \right)^{1/4}.$$

The first is also called the $L_p$ norm, and the second is also called the Gowers $U_2$ norm, or the *additive energy* of $f$. The same definitions hold for $\hat{f}$ except we instead normalise by taking sums rather than expectations. Finally, we have the following two equalities:

$$||f||_2 = ||\hat{f}||_2 \quad \textit{(Parseval's Theorem)}$$
$$||\hat{f}||_{U_2} = ||f||_4. \qquad (\star)$$

In this thesis, we will consider the ratio between different $L_p$ norms of functions on $\mathbb{F}_2^n$. Bounds on ratios between norms of functions of this form are closely linked to the Hypercontractivity Theorem, also called the Bonami-Beckner inequality [7, 3], which bounds the ratio between the $L_p$ norm of a function and the $L_q$ norm of a "smoothing" of it, for certain $q \geqslant p$. Also, bounds on ratios between norms of functions of the form above are sometimes called *Khintchine-type inequalities*, after the prototypical Khintchine inequality which bounds the ratio $||f||_2/||f||_1$ [38].

The quantity $||f||_p/||f||_2$ has been investigated by Kirshner and Samorodnitsky [39, 40]. In particular, they investigate the maximal value of this ratio among functions $f$ for which the support of $\hat{f}$ has been restricted to a Hamming sphere. The functions with Fourier support restricted to the Hamming sphere of radius $k$ are precisely the homogeneous polynomials of degree $k$, making this a fairly natural consideration.

In [39], Kirshner and Samorodnitsky focus on the ratio $||f||_4/||f||_2$ among such functions. The ratio between the $L_4$ and $L_2$ norms is a fairly natural one to consider: in view of the relation $(\star)$ above, this can be phrased as a statement about the additive energy of a function $f$. Kirshner and Samorodnitsky conjecture that this ratio is maximised for the function whose Fourier transform is constant on the Hamming sphere; this function is also called the *Krawchouk Polynomial*. They also prove that this conjecture is true to leading order. In Chapter 5 of this thesis, we will prove their conjecture.

Observe that we have $||f||_4/||f||_2 = ||\hat{f}||_{U_2}/||\hat{f}||_2$, as a consequence of the properties of $f$ and $\hat{f}$ above. Thus, by replacing $f$ with $\hat{f}$ it will suffice to show that among functions $f$ supported on the Hamming sphere, $||f||_{U_2}/||f||_2$ is maximised for the constant function.

Finally, we will consider a problem related to the Polynomial Freiman-Ruzsa conjecture over $\mathbb{F}_p^n$. Recall the statement of Freiman's theorem [24].

**Theorem** (Freiman's Theorem)**.** *Let $p$ be a prime. There exists a function $f : \mathbb{N} \to \mathbb{N}$ with the following property.*

*Suppose that $K$ is a constant. If $A \subseteq \mathbb{F}_p^n$ is a set satisfying $|A + A| < K|A|$, then $A$ can be covered by $f(K)$ cosets of a subspace of size at most $|A|$.*

Ruzsa's original proof [49] gives $f(K)$ exponential in $K^{O(1)}$ .

A useful equivalent form is as follows (see [24, Proposition 10.2]).

**Theorem** (Freiman's Theorem, equivalent form)**.** *Let $p$ be a prime. There exists a function $f_1 : \mathbb{N} \to \mathbb{N}$ with the following property.*

*Suppose that $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$ is a function and let $S = \{\phi(x+y) - \phi(x) - \phi(y)\}$.*

*Then there is a linear map $\tilde{\phi} : \mathbb{F}_p^n \to \mathbb{F}_p^N$ with the property that $\phi - \tilde{\phi}$ takes at most $f_1(|S|)$ values.*

The general idea behind the usefulness of Freiman's theorem is that it provides a link between a set having additive structure (having small doubling), and algebraic structure (being 'close' to a subgroup). Naturally, any set $A$ which is contained in a subspace of size at most $K'|A|$ has doubling at most $K'$, because a subspace has doubling 1. Therefore, we can write the following weak equivalence:

$$\begin{gathered}\text{doubling} \leqslant K \\ \Downarrow \\ \text{covered by } f(K) \text{ cosets of a subspace of size} \leqslant |A| \\ \Downarrow \\ \text{doubling} \leqslant f(K)^2.\end{gathered}$$

With this in mind, it is natural to want good bounds on $f$. Indeed, if we had good bounds on $f$, then we would be able to naturally pass between good 'additive' structure and good 'algebraic' structure with only small losses. The widely believed Polynomial Freiman-Ruzsa Conjecture, first proposed by Marton, asserts the following:

**Conjecture** (PFR Conjecture)**.** *The function $f(K)$ in the statement of Freiman's Theorem may be taken to be a polynomial in $K$.*

This is equivalent to the assertion that the function $f_1$ in the second form of Freiman's Theorem may be taken to be a polynomial. This form of the conjecture, concerning a function $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$, is particularly appealing. The following strengthened form has been considered [53].

**Conjecture** (Strengthened PFR)**.** *There is a constant $t$ such that the following holds. Suppose that $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$ is a function and let $S = \{\phi(x+y) - \phi(x) - \phi(y)\}$.*

*Then there is a linear map $\tilde{\phi}$ with the property that $\phi - \tilde{\phi}$ takes values in the set $tS$.*

Observe that this implies the Polynomial Freiman-Ruzsa conjecture, if we take $f_1(x) = x^t$.

In [53], Green and Tao present a counterexample to this conjecture. However, their example only works over $\mathbb{F}_2$, as it relies on the fact that quadratic forms over $\mathbb{F}_2$ need not be diagonalisable. In Chapter 6 of this thesis, we will discuss a different example which disproves this conjecture, and which works in arbitrary characteristic.

# Notation

As is standard, we will use the notation $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ to denote the positive integers, integers, rationals and real numbers respectively. For a prime power $q$, we will let $\mathbb{F}_q$ denote the finite field on $q$ elements.

Throughout this thesis, we will use the Vinogradov notation $f \ll g$ to denote that $f = O(g)$. When the $\ll$ is subscripted, we will allow the implicit constant to depend on the subscripts.

We will also use the standard additive combinatorial notation regarding set operations. In particular,

$$\begin{aligned} A + B &= \{a + b : a \in A, b \in B\} \\ A - B &= \{a - b : a \in A, b \in B\} \\ rA &= \underbrace{A + \cdots + A}_{r \text{ copies of } A} \\ r \cdot A &= \{ra : a \in A\}. \end{aligned}$$

We will also use the notation $[n] = \{1, 2, \ldots, n\}$, and $[m, n] = \{m, m+1, \ldots, n\}$.

# Chapter 2

# A connection between matchings and removal in abelian groups

## 2.1 Introduction

In an abelian group $G$, define a *triangle* to be a triple of elements $x$, $y$ and $z$ with $x + y + z = 0$. Green's arithmetic triangle removal lemma [25] states that for any $\varepsilon > 0$, there is a $\delta > 0$ such that the following holds. Whenever $X$, $Y$ and $Z$ are subsets of $G$ such that there are at most $\delta N^2$ triangles $x + y + z = 0$ with $x \in X$, $y \in Y$ and $z \in Z$, we can remove at most $\varepsilon N$ elements from $X$, $Y$ and $Z$ to remove all of the triangles. The bounds in [25] are quite weak; $1/\delta$ is given as a tower of twos of height polynomial in $1/\varepsilon$. The best known bounds for this problem in general are still of tower type.

In [20], Fox and Lovász proved a much stronger bound on $\delta$ in the case of $G = \mathbb{F}_p^n$ for a fixed prime $p$; namely, that $1/\delta$ is bounded by a polynomial in $1/\varepsilon$. Define an *additive matching* to be a collection of triples $(x_i, y_i, z_i)$ such that $x_i + y_j + z_k = 0$ if and only if $i = j = k$. These are also called tricolored sum-free sets, and can be represented by $(X, Y, Z)$, where $X = \{x_i\}$, $Y = \{y_i\}$ and $Z = \{z_i\}$. Building on the groundbreaking work on the cap set problem by Croot-Lev-Pach [15] and subsequent work by Ellenberg-Gijswijt [18], Kleinberg [41] gave a polynomial upper bound for the size of an additive matching in $G$ in the case that $G = \mathbb{F}_2^n$, and Blasiak-Church-Cohn-Grochow-Naslund-Sawin-Umans [6] extended this to $\mathbb{F}_q^n$ for a fixed prime power $q$. The argument by Fox and Lovász made use of these results to prove the polynomial bounds on removal.

Polynomial bounds on removal are much stronger than could possibly hold in general groups. Indeed, using Behrend's construction [4] of a large subset of $\mathbb{Z}/N\mathbb{Z}$

with no three-term arithmetic progressions, it is possible [50] to show that the best one could hope for is

$$\varepsilon \ll \exp\left(-c\sqrt{\log(1/\delta)}\right).$$

The goal of this section is to adapt the arguments of Fox and Lovász to show that, in the context of cyclic groups, good bounds on additive matchings give good bounds on removal.

Assume that, in a cyclic group of order $M$, the density of an additive matching is bounded above by $f(M)$ for some function $f$. Assume that $f(M)$ can be taken to be decreasing as $M$ increases, but that $Af(A) < Bf(B)$ for $A < B$; these conditions correspond to the claim that the maximum size of an additive matching increases as the size of the group increases, but the maximum density decreases. Observe that Behrend's example guarantees that

$$f(M) \gg \exp\left(-c\sqrt{\log M}\right)$$

because, if $A$ is a progression free set, then $(A, -2A, A)$ is an additive matching.

We will also need a function $g$ such that

- $g(\rho)$ increases as $\rho$ decreases,
- $\sum_{i=1}^{\infty} \frac{1}{g(2^{-i})} \leqslant \frac{1}{4}$,
- For some absolute constant $\alpha$, $g(\rho)f\left(\frac{1}{g(\rho)\rho}\right)$ is decreasing as $\rho$ decreases, for $\rho < \alpha$.

Here, $g$ is just a slowly-growing function used in our argument. It is not especially important what $g$ is; typically, we will take $g$ to be a function of the form $g(\rho) = \log^{1+c}\rho$ for a constant $c$, to ensure that the second property holds. The third condition will then put a constraint on the growth of $f$ for our arguments to work. In particular, if $f(N)$ is as large as $\log^{-2} N$, then we will not be able to find a function $g$ satisfying the conditions.

We are now ready to state Theorem 2.1.1.

**Theorem 2.1.1.** *Assume that a function $g$ satisfying the above conditions exists.*

*Suppose that $A$, $B$ and $C$ are subsets of $\mathbb{Z}/N\mathbb{Z}$ for some $N \in \mathbb{N}$ with the property that there are at most $\delta N^2$ triangles $a + b + c = 0$ with $a \in A$, $b \in B$ and $c \in C$.*

*Then, we can remove all of the triangles by deleting at most $\varepsilon N$ elements from $A$, $B$ and $C$, where $\varepsilon$ satisfies*

$$\varepsilon \ll \sqrt{g(\delta) f\left(\frac{1}{g(\delta)\delta}\right)}. \tag{2.1.1}$$

We can deduce some consequences of this:

**Corollary 2.1.2.** *Suppose that we have the best possible bound on the size of an additive matching, namely a Behrend-type bound. In particular, we can take $f(N)$ to be $\exp(-c\sqrt{\log N})$ for some constant $c$. Then $g(\rho) = k\log^2(1/\rho)$ suffices, and we deduce the bound*

$$\varepsilon \ll \exp(-c_1\sqrt{\log 1/\delta})$$

*for some other constant $c_1$.*

**Corollary 2.1.3.** *Suppose that the much more pessimistic $f(N) = \log^{-2-\gamma} N$ holds, for some constant $\gamma > 0$. Then, we can take $g(\rho) = k\log^{1+\gamma/3}(1/\rho)$, and we deduce that*

$$\varepsilon \ll \log(1/\delta)^{-O(1)}.$$

*Remark.* A converse of sorts to Theorem 2.1.1, namely that bounds on removal imply similar bounds on the maximal size of an additive matching, is relatively trivial. Indeed, suppose that, whenever subsets $X$, $Y$ and $Z$ of a cyclic group $G = \mathbb{Z}/N\mathbb{Z}$ define at most $\delta N^2$ triangles, the triangles can be removed by deleting at most $\varepsilon N$ elements, where $\varepsilon \ll f(1/\delta)$.

Then, an additive matching of size $\theta N$ defines at most $N = \frac{1}{N}N^2$ triangles, and requires removal of at least $\theta N$ elements to remove the triangles. Thus, $\theta \ll f(N)$, which is the converse we wanted.

## 2.2 Theorem 2.1.1 for $N$ prime

In this section, following the approach in [20], we prove Theorem 2.1.1, in the case that $N$ is prime. We start with a lemma, which is an analogue of Lemma 5 from [20]:

**Lemma 2.2.1.** *Suppose we have three subsets of $\mathbb{Z}/N\mathbb{Z}$, $X$, $Y$ and $Z$, with $\delta N^2$ triangles $(x, y, z) \in X \times Y \times Z$. Suppose further that, for each $x \in X$, there are at most $\rho N$ triangles $(x, y', z') \in X \times Y \times Z$, and the same for $X, Y$ and $Z$ permuted.*

*Then, we deduce that*

$$\delta \leqslant 250\rho f(\rho^{-1}). \tag{2.2.1}$$

*Proof.* Choose $a$, $b$ and $d$ uniformly and independently from $\mathbb{Z}/N\mathbb{Z}$, such that $d$ is nonzero. Let $L = 2l + 1$ be an odd positive integer such that $1/20 \leqslant L\rho \leqslant 1/10$. We may assume that $\rho$ is small enough that $L > 20$ by adjusting the implicit constant in 2.2.1, and so such a choice of $L$ exists.

We say that a triangle $x + y + z = 0$ is *valid* if and only if

$$\begin{aligned} x \in I_X &:= a + [-l, l]d \\ y \in I_Y &:= b + [-l, l]d \\ z \in I_Z &:= -a - b + [-2l, 2l]d. \end{aligned}$$

We say that a valid triangle $x + y + z = 0$ is *good* provided that each of $x$, $y$ and $z$ is in only one valid triangle, namely $(x, y, z)$. Observe that a triangle $(x, y, z)$ is good if and only if $x \in I_X$ and $y \in I_Y$, as those two imply that $z \in I_Z$.

**Claim 1.** *Given a valid triangle $x + y + z = 0$, it has a probability at least $2/5$ of being good.*

*Proof of Claim 1.* We first show that the probability that $x$ is in another valid triangle is at most $1/5$. For each $y'$ that forms a triangle with $x$, it has a probability of $\frac{L-1}{N-1} \leqslant \frac{L}{N}$ of lying in $I_Y$, because choosing values $a$, $b$ and $d$ such that $x + y + z = 0$ is valid is equivalent to choosing values $r, s \in [-l, l]$ so that $x = a + rd, y = b + sd$, and then choosing any value of $d$. Thus, each value of $y' \neq y$ will occur with probability $\frac{L-1}{N-1}$ since, for each choice of $t \neq s \in [-l, l]$, there is exactly one choice of $d$ such that $y = b + td$.

There are at most $\rho N$ possible choices of $y'$ to consider, so the union bound guarantees that the probability that $x$ is in another valid triangle is at most $L\rho \leqslant 1/5$.

The same argument applies to the probability that $y$ or $z$ is in another valid triangle. Thus, the probability that either $x$, $y$ or $z$ cause the triangle to be not good is at most $3/5$ by the union bound, and thus the probability that the triangle is good is at least $2/5$. ◊

**Claim 2.** *The expected number of good triangles is at least $\frac{\delta}{250\rho^2}$.*

*Proof of Claim 2.* There are $\delta N^2$ triangles. The probability that any such a triangle is valid is exactly $\frac{L^2}{N^2} \geqslant \frac{1}{100\rho^2 N^2}$, and the probability that a valid triangle is good is at least $\frac{2}{5}$.

Thus, the probability that a triangle is good is at least $\frac{2}{500\rho^2 N^2}$, and so the expected number of good triangles is at least $\frac{\delta}{250\rho^2}$. ◊

Thus, there must exist a choice of parameters $a, b$ and $d$ for which there are at least $\frac{\delta}{250\rho^2}$ good triangles.

Now, observe that we may map the intervals $I_X$, $I_Y$ and $I_Z$ into $\mathbb{Z}/M\mathbb{Z}$, where $M = \rho^{-1}$, in the obvious way. For example, $a + rd \in I_X$ for $r \in [-l, l]$ maps to $r$ mod $M$. It is clear that this map sends triangles to triangles; since $M > 8l$, this map also preserves the status of not being a triangle.

In other words, the at least $\frac{\delta}{250\rho^2}$ good triangles we found correspond to an additive matching within $\mathbb{Z}/M\mathbb{Z}$. Given our hypothesis that the size of an additive matching is bounded above by $Mf(M)$, we deduce that

$$\frac{\delta}{250\rho^2} \leqslant Mf(M) \tag{2.2.2}$$

and thus

$$\delta \leqslant 250\rho f(\rho^{-1}) \tag{2.2.3}$$

which is exactly what we sought. □

Next, we prove an analogue of Lemma 6 from [20].

**Lemma 2.2.2.** *Suppose that $\varepsilon, \delta > 0$ satisfy*

$$\varepsilon \gg g(\delta)\sqrt{f\left(\frac{1}{g(\delta)\delta}\right)}$$

*for the functions $f$ and $g$ defined previously, and that $\delta < \alpha$ (which was defined immediately before Theorem 2.1.1).*

*Suppose we have a collection of $\varepsilon N$ disjoint triangles $x_i + y_i + z_i = 0$, and let $X = \{x_i\}$, defining $Y$ and $Z$ analogously. Then there must be at least $\delta N^2$ triangles $x_i + y_j + z_k = 0$.*

*Proof.* Suppose we have fewer than $\delta N^2$ triangles. We will remove points from $X, Y$ and $Z$ until there are no points contained in significantly more triangles than average.

Let us begin a process defined as follows. If, at some stage, we have $\delta' N^2$ triangles, we will remove some point in more than $g(\delta')\frac{\delta'}{\varepsilon}N$ triangles. We will continue until this is no longer possible.

During the period where the number of triangles is between $\beta N^2$ and $2\beta N^2$, each removal will remove at least $g(\beta)\frac{\beta}{\varepsilon}N$ triangles. This can happen at most $\frac{2\varepsilon}{g(\beta)}N$ times

because there are at most $2\beta N^2$ triangles to remove. Thus, we will remove in total at most

$$\frac{2\varepsilon}{g(\delta)}N + \frac{2\varepsilon}{g(\delta/2)}N + \ldots \leqslant \frac{\varepsilon}{2}N$$

points. In particular, $X, Y$ and $Z$ will each be left with at least $\frac{\varepsilon}{2}N$ elements. And, for some $\delta' \leqslant \delta$, we have $\delta' N^2$ triangles, and each point is contained in at most $g(\delta')\frac{\delta'}{\varepsilon}N$ triangles.

Now, we will apply Lemma 2.2.1. We set $\delta'$ as $\delta$, and $\rho = g(\delta')\frac{\delta'}{\varepsilon}$. This gives us

$$\begin{aligned}
\delta' &\leqslant 250 g(\delta')\frac{\delta'}{\varepsilon} f\left(\frac{\varepsilon}{g(\delta')\delta'}\right) \\
\varepsilon &\leqslant 250 g(\delta')\varepsilon^{-1} f\left(\frac{1}{g(\delta')\delta'}\right) \\
\varepsilon^2 &\leqslant 250 g(\delta) f\left(\frac{1}{g(\delta)\delta}\right)
\end{aligned}$$

where in the second line we used the conditions on $f$ and in the third line we used that the right hand side decreases as $\delta'$ decreases. □

We are now ready to prove Theorem 2.1.1.

*Proof of Theorem 2.1.1.* We follow the same strategy as in [20]. Suppose that $A$, $B$ and $C$ are such that we cannot remove all the triangles without removing at least $\varepsilon N$ elements from $A$, $B$ and $C$. Any maximal set of disjoint triangles must have size at least $\frac{\varepsilon}{3}N$, else we could remove all of the elements of those triangles and there would be no triangles left.

Lemma 2.2.2 guarantees that we must have at least $\delta N^2$ triangles in total, where

$$\varepsilon \ll \sqrt{g(\delta) f\left(\frac{1}{g(\delta)\delta}\right)}$$

as required. □

## 2.3 Theorem 2.1.1 for general $N$

In this section, we complete the proof of Theorem 2.1.1 in the case that $N$ need not be prime. First, observe that the prime case of Theorem 2.1.1 implies that it holds for subsets of the interval $[-M/2, M/2]$. Indeed, provided that the functions $f$ and $g$ exist and satisfy all of the requirements imposed upon them above, then we can deduce the following:

**Corollary 2.3.1.** *Suppose that $A$, $B$ and $C$ are subsets of $[-M/2, M/2]$ with the property that there are at most $\delta M^2$ triangles $a+b+c=0$ with $a \in A$, $b \in B$ and $c \in C$.*

*Then, we can remove all of the triangles by deleting at most $\varepsilon M$ elements from $A$, $B$ and $C$, where $\varepsilon$ satisfies*

$$\varepsilon \ll \sqrt{g(\delta) f\left(\frac{1}{g(\delta)\delta}\right)}. \tag{2.3.1}$$

*Proof.* Suppose we have sets $A$, $B$ and $C$ which define $\delta M^2$ triangles. Select a prime $N$ such that $2M \leqslant N \leqslant 4M$, and consider the reduction modulo $N$ map $\phi$ taking $[-M/2, M/2]$ to $\mathbb{Z}/N\mathbb{Z}$. This preserves the status of being a triangle, as well as the status of not being a triangle.

The image of $(A, B, C)$ under $\phi$ contains at most $\delta(N/2)^2$ triangles, and thus requires removal of at most $\varepsilon N$ points to remove all of the triangles, where $\varepsilon$ satisfies

$$\varepsilon \ll \sqrt{g(\delta/4) f\left(\frac{4}{g(\delta/4)\delta}\right)}. \tag{2.3.2}$$

Thus, to remove the triangles from $(A, B, C)$, the deletion of at most $\varepsilon N \leqslant 4\varepsilon M$ points is necessary. By adjusting the implicit constant, we deduce Corollary 2.3.1. □

We may now deduce that Theorem 2.1.1 holds in arbitrary finite cyclic groups:

**Corollary 2.3.2.** *Theorem 2.1.1 holds without the requirement that $N$ is prime.*

*Proof.* Suppose not; then for some (composite) $N$, $\mathbb{Z}/N\mathbb{Z}$ contains sets $A$, $B$ and $C$ which define at most $\delta N^2$ triangles, but require deletion of at least $\varepsilon N$ points to remove the triangles, and where $\varepsilon$ does not satisfy (2.1.1) (with a slightly adjusted implicit constant).

As in the proof of Theorem 2.1.1, a greedy argument guarantees the existence of $\frac{\varepsilon}{3}N$ disjoint triangles. Consider the map $\pi$ taking $\mathbb{Z}/N\mathbb{Z}$ to $[0, N-1]$ in the obvious way; a triple $(a, b, c)$ in $\mathbb{Z}/N\mathbb{Z}$ is a triangle if and only if its image under $\pi$ has sum either $N$ or $2N$.

For one of the two possibilities for the sum, there are at least $\frac{\varepsilon}{6}N$ disjoint triangles. In the first case, in which the triangles have sum $N$ in the image of $\pi$, consider $\pi(A)$, $\pi(B)$ and $\pi(C) - N$, and in the second case, consider $\pi(A)$, $\pi(B) - N$ and $\pi(C) - N$. Either way, we have at most $\frac{\delta}{2}(2N)$ triangles in $[-N, N]$, which require deletion of at least $\frac{\varepsilon}{12}(2N)$ points to remove, because they define at least that many disjoint triangles. Corollary 2.3.1 gives us the result. □

# Chapter 3

# Maximising the number of solutions to a linear equation in a set of integers

## 3.1 Introduction

Let $a_1, a_2$ and $a_3$ be fixed coprime integers, none of which is zero. We will consider the linear equation

$$a_1x_1 + a_2x_2 + a_3x_3 = 0. \tag{3.1.1}$$

In this section, we are interested in the problem of finding sets with as many solutions to (3.1.1) as possible. This leads to the following definition.

**Definition 3.1.1.** Given a finite set $S \subseteq \mathbb{Z}$, define $T(S) = T_{a_1,a_2,a_3}(S)$ to be the number of triples $x_1, x_2, x_3 \in S$ satisfying (3.1.1).

The trivial upper bound on $T(S)$ is $T(S) \leqslant |S|^2$. This is because, for any choice of $x_1$ and $x_2$, there is at most one choice of $x_3$ such that $a_1x_1 + a_2x_2 + a_3x_3 = 0$, namely $x_3 = \frac{-a_1x_1-a_2x_2}{a_3}$. We are interested in making $T(S)$ as large as possible, for a fixed size $|S|$.

For some choices of coefficients $a_1, a_2$ and $a_3$, the exact maximal value of $T(S)$ is known. For example, consider the case $a_1 = a_2 = a_3 = 1$. Then, work of Hardy and Littlewood [32] and Gabriel [21] shows that, when $|S|$ is odd, $T(S)$ is maximised when $S$ is an interval centred about 0. This was extended to even $|S|$ by Lev in [43]. In fact, their arguments show that if $S \subseteq \mathbb{Z}$ is a set, and $S'$ is an interval centred about 0 of the same size, then $T_{a_1,a_2,a_3}(S) \leqslant T_{1,1,1}(S')$ for any $a_1, a_2$ and $a_3$. The ideas behind their approaches involve rearrangement inequalites, which are discussed in detail in [33, Chapter 10], and which inspire some of the arguments we use in this section.

Similarly, it is shown by Green and Sisask in [27, Theorem 1.2] and by Lev and Pinchasi in [44, Theorem 2] respectively that, if $(a_1, a_2, a_3) = (1, -2, \pm 1)$, then $T(S)$ is again maximised when $S$ is an interval centred at 0.

The set of solutions to $x_1 - 2x_2 + x_3 = 0$ is precisely the set of three-term arithmetic progressions; that is, the set of affine shifts of the set $\{0, 1, 2\}$. By analogy with this, Bhattacharya, Ganguly, Shao and Zhao considered longer arithmetic progressions; in [5, Theorem 2.4], they proved that the number of $k$ term arithmetic progressions in a set $S$ of $n$ integers is maximised when $S$ is an interval.

Ganguly asked [22] about other affine patterns; in particular, finding sets $S$ with as many affine copies of $\{0, 1, 3\}$, or solutions to $x + 2y = 3z$, as possible. In this case, such a result would necessarily be less clean; for instance, there are more solutions to $x + 2y = 3z$ in $\{0, 1, 3\}$ than in $\{0, 1, 2\}$.

Indeed, in general, much less is known. For a lower bound on the maximal value of $T(S)$, a fairly good bound is given by the following example.

**Proposition 3.1.2.** *Regardless of the values of $a_1, a_2$ and $a_3$, there are choices of $S$ with $|S|$ arbitrarily large, for which $T(S) \geqslant \frac{1}{12}|S|^2 + O(|S|)$.*

As far as the author is aware, this construction does not explicitly appear in the literature. However, it is a fairly natural construction, and it seems probable that it has been considered before.

*Proof.* The idea behind the construction is to split $S$ into three pieces $S_1, S_2$ and $S_3$, of roughly equal size, for which there are many solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$ with each $x_i$ taken from $S_i$. Let $M$ be a large integer, which we assume to be divisible by 6. We will define

$$\begin{aligned} S_1 &= a_2a_3[-M/6, M/6] \\ S_2 &= a_1a_3[-M/6, M/6] \\ S_3 &= a_1a_2[-M/6, M/6], \end{aligned}$$

where $[-M/6, M/6]$ refers to the set of integers with absolute value no greater than $M/6$, and set $S = S_1 \cup S_2 \cup S_3$. Then, $|S|$ is certainly no more than $M$.

However, we may find a large collection of triples $(x_1, x_2, x_3)$ by choosing $x_1 \in S_1$ and $x_2 \in S_2$ arbitrarily, and selecting those for which $x_3 = \frac{-a_1x_1 - a_2x_2}{a_3}$ is in $S_3$. If $x_1 = a_2a_3x_1'$ and $x_2 = a_1a_3x_2'$, then we have $x_3 = -a_1a_2(x_1' + x_2')$. Therefore, a pair $(x_1', x_2')$ will give rise to a solution precisely when $|x_1' + x_2'| \leqslant M/6$.

We may compute the number of such pairs $(x_1', x_2')$ as the sum

$$\sum_{x_1'=-M/6}^{M/6} M/3 + 1 - |x_1'| = \frac{1}{12}M^2 + O(M).$$

Thus, the number of triples is at least $\frac{1}{12}|S|^2 + O(|S|)$. □

Given this, it is natural to define the following quantity:

**Definition 3.1.3.** Define $\gamma_{a_1,a_2,a_3}$ by

$$\gamma_{a_1,a_2,a_3} = \limsup_{|S|\to\infty} \frac{T(S)}{|S|^2}$$

where $S$ runs over subsets of $\mathbb{Z}$.

Thus, the assertion that

$$\frac{1}{12} \leqslant \gamma_{a_1,a_2,a_3} \leqslant \frac{3}{4} \tag{3.1.2}$$

holds for all $a_1, a_2$ and $a_3$ follows from Proposition 3.1.2 and the work of Hardy and Littlewood in [32].

*Remark.* The lim sup in the definition of $\gamma_{a_1,a_2,a_3}$ refers to the limit supremum as $S$ runs over all sets of integers. This definition could be rewritten as

$$\gamma_{a_1,a_2,a_3} = \limsup_{n\to\infty} \max_{|S|=n} \frac{T(S)}{|S|^2}.$$

It is a natural to ask whether the lim sup in this formulation may be replaced by a limit; in other words, whether or not $\max_{|S|=n} \frac{T(S)}{|S|^2}$ converges as $n$ tends to infinity. Questions of this type have been considered before; for instance, a result of a similar flavour has been proven by Croot in [14].

We will see later that, indeed, the answer to this question is yes. In other words, we could replace the lim sup with a lim in this second formulation of the definition of $\gamma_{a_1,a_2,a_3}$.

Exact values for $\gamma_{a_1,a_2,a_3}$ are only known in cases for which $|a_1a_2a_3| \leqslant 2$ (this includes the cases previously discussed). In particular, we have

$$\gamma_{1,1,\pm1} = \frac{3}{4} \tag{3.1.3}$$

$$\gamma_{1,-2,\pm1} = \frac{1}{2} \tag{3.1.4}$$

$\gamma_{1,-2,1}$ is [27, Theorem 1.2], and $\gamma_{1,-2,-1} = \frac{1}{2}$ is [44, Theorem 2]. The same holds in the third non-equivalent case with $|a_1a_2a_3| = 2$, namely $\gamma_{1,2,1} = \frac{1}{2}$, and we will

prove this in Section 3.8. Even the value of $\gamma_{1,2,-3}$ is not known, although the author conjectures that it is $\frac{1}{3}$, which is the value calculated for $S = [-M/2, M/2]$.

The main theorem of this section is a converse, of sorts, to Proposition 3.1.2. In particular, we will prove the following.

**Theorem 3.1.4.** *The constant $\frac{1}{12}$ in the statement of Proposition 3.1.2 is optimal, in the following sense. For any $\varepsilon > 0$, there exists a choice of $a_1, a_2$ and $a_3$ for which $\gamma_{a_1,a_2,a_3} \leqslant \frac{1}{12} + \varepsilon$.*

In view of this theorem, (3.1.2) gives the best possible bounds on $\gamma_{a_1,a_2,a_3}$ that are independent of the coefficients $a_i$.

In this chapter, we will present two approaches to prove Theorem 3.1.4. The first approach is elementary, and we will present it in Section 3.2. The second relies on the arithmetic regularity lemma, but as an intermediate lemma proves the following.

**Lemma 3.1.5.** *To compute $\gamma_{a_1,a_2,a_3}$, it suffices to restrict our attention to sets of the following form. Choose an integer $q \in \mathbb{N}$, with the property that if $p$ is a prime factor of $q$, then $p$ divides some $a_i$. Then, choose a function $f : \mathbb{Z}/q\mathbb{Z} \to [0,1]$ with the property that $\sum_{x \in \mathbb{Z}/q\mathbb{Z}} f(x) = 1$.*

*Then, for any large $M \in \mathbb{N}$, set $S = S_M$ to be*

$$S_M = \bigcup_{x \in \mathbb{Z}/q\mathbb{Z}} (\tilde{x} + q[-Mf(x)/2, Mf(x)/2]), \tag{3.1.5}$$

*where $\tilde{x}$ is the representative of $x$ in $[0, q-1]$.*

For example, consider the equation $x + y = z$, which corresponds to the problem of calculating $\gamma_{1,1,-1}$. The only choice of $q$, all of whose prime factors divide one of the coefficients, is $q = 1$, so Lemma 3.1.5 says that to calculate $\gamma_{1,1,-1}$ it suffices to consider sets of the form $[-M/2, M/2]$. From this, it is trivial to prove that $\gamma_{1,1,-1} = 3/4$.

In most cases, it will be substantially more difficult to determine $\gamma_{a_1,a_2,a_3}$, even armed with Lemma 3.1.5. However, in Section 3.8 we will present an argument that uses it to prove $\gamma_{1,2,1} = 1/2$, which is the case of $|a_1a_2a_3| = 2$ not covered by the results of Green-Sisask [27] or Lev-Pinchasi [44].

We will prove Lemma 3.1.5 in Section 3.3, and then show how it can be used to prove Theorem 3.1.4 in Section 3.4. We will also prove the claim made in the remark after Definition 3.1.3 about replacing lim sup with lim in Section 3.3.

One might also ask about generalising Theorem 3.1.4 to other settings. For instance, we may wish to consider equations in $k$ variables for any $k \geqslant 3$. It turns out that in this setting, an analogue of Proposition 3.1.2 holds. Set

$$\sigma_k = \int_{-\infty}^{\infty} \left( \frac{\sin \pi x}{\pi x} \right)^k dx, \tag{3.1.6}$$

so, for example, $\sigma_3 = 3/4$. Then, for any choice of coefficients $a_1, \dots, a_k$, there are sets $S$ with at least $\frac{\sigma_k}{k^{k-1}}|S|^{k-1} + O(|S|^{k-2})$ solutions to $a_1x_1 + \cdots + a_kx_k = 0$.

Furthermore, the corresponding analogue of Theorem 3.1.4 holds. For any $\varepsilon > 0$, there are choices of coefficients $a_1, \dots, a_k$ for which any large set $S$ must have no more than $\left(\frac{\sigma_k}{k^{k-1}} + \varepsilon\right)|S|^{k-1}$ solutions. For instance, for any small positive $\varepsilon$ we can find coefficients $a_1, a_2, a_3$ and $a_4$ with the property that $T(S) \leqslant \left(\frac{1}{96} + \varepsilon\right)|S|^3$, where $T(S)$ counts the number of solutions to $a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 = 0$.

We have two proofs of Theorem 3.1.4, and analogously there are two ways to deal with the $k$ variable case. One way to do this would involve adapting the elementary proof of Theorem 3.1.4; we will do this in Section 3.5. Alternatively, we could prove a generalisation of Lemma 3.1.5 and use that to generalise Theorem 3.1.4; we will do so in Section 3.6.

Alternatively, we may wish to consider systems of more than one linear equation. In that case, the opposite is true. Indeed, it is possible to show that there is *no* constant $c > 0$, such that for any system of 2 equations in 4 variables, there are large sets $S$ with at least $c|S|^2$ solutions to the system. We will prove this fact in Section 3.7.

Finally, we will prove some further results related to calculating $\gamma_{a_1,a_2,a_3}$. First, we will demonstrate how Lemma 3.1.5 may be used to prove that $\gamma_{1,2,1} = 1/2$ in Section 3.8. We will then discuss the family of equations of the form $x + y = kz$. In particular, we will show that, if we restrict to this family, analogues of Proposition 3.1.2 and Theorem 3.1.4 hold with the constant $\frac{1}{12}$ replaced by $\frac{1}{5}$. Furthermore, a careful inspection of our argument will yield some results about the structure of the set of numbers of the form $\gamma_{a_1,a_2,a_3}$ where $a_1, a_2$ and $a_3$ are any nonzero coprime integers. We will do this in Section 3.9.

As we have already noted, $T(S)$ will be the number of solutions to the equation (3.1.1) in $S$. If $T$ is subscripted, the subscripts refer to fixed coefficients; for instance, $T_{1,1,1}(S)$ is the number of solutions to $x_1 + x_2 + x_3 = 0$ in $S$. We can extend this definition in a number of ways.

- For sets $S_1, S_2$ and $S_3$, let $T_{a_1,a_2,a_3}(S_1, S_2, S_3) = T(S_1, S_2, S_3)$ be the number of solutions to (3.1.1) where $x_i$ is taken from $S_i$.

- For a function $f : \mathbb{Z} \to [0, 1]$ with finite support, define

$$T(f) = \sum_{a_1x_1+a_2x_2+a_3x_3=0} f(x_1)f(x_2)f(x_3),$$

  observing that this sum is finite. This reduces to the original definition when we set $f$ to be the indicator function $\mathbb{1}_S$ of $S$.

Let us also define $t(S) = T(S)/|S|^2$, so $\gamma_{a_1,a_2,a_3} = \limsup_{|S|\to\infty} t(S)$. We may define $t(f)$ analogously.

We will also frequently refer to intervals of $\mathbb{R}$. Let $I_a$ denote an interval of length $a$ centred at 0. Thus, $I_a = [-a/2, a/2]$. We will also use this to denote the indicator function, so $I_a(x)$ is 1 if $|x| \leqslant a/2$, and 0 otherwise.

## 3.2 The elementary approach to Theorem 3.1.4

In this section, we will give the elementary proof of Theorem 3.1.4, which avoids relying on the arithmetic regularity lemma. Before we proceed with the proof, we will present some additive combinatorial results that will be necessary for this proof.

### 3.2.1 Additive combinatorial lemmas

For any set $A \subseteq \mathbb{Z}$, let $\delta[A]$ be its growth under the differencing operator, $\frac{|A-A|}{|A|}$. If $A$ and $B$ are two sets of integers, let the *additive energy* between $A$ and $B$, $E(A, B)$, be defined by

$$E(A, B) = \#\{(a_1, b_1, a_2, b_2) \in A \times B \times A \times B : a_1 + b_1 = a_2 + b_2\}.$$

It is easy to see that this satisfies the following inequalities:

$$\begin{aligned} E(A, B) &\leqslant |A|^2|B| \\ E(A, B) &\leqslant |A||B|^2 \\ E(A, B) &\leqslant |A|^{3/2}|B|^{3/2}. \end{aligned} \tag{3.2.1}$$

To see why, observe that the first equation holds because for each triple $(a_1, a_2, b_1)$ there can be at most one choice of $b_2 \in B$ with $a_1 + b_1 = a_2 + b_2$, the second holds by symmetry and the third follows by multiplying the first two.

We will require the following lemma, which states that, when two sets $A$ and $B$ are such that $\delta[A]$ and $\delta[B]$ are small, and if $E(A, B)$ is large, then $|A - B|$ is also small.

**Lemma 3.2.1** ([27, Lemma 3.1 (iv)]). *Suppose that $A, B \subseteq \mathbb{Z}$ are sets with $E(A,B) \geqslant \eta|A|^{3/2}|B|^{3/2}$.*

*Then, $|A-B| \leqslant \frac{\delta[A]\delta[B]}{\eta}|A|^{1/2}|B|^{1/2}$.*

We will also require a structure theorem due to Green and Sisask.

**Theorem 3.2.2** ([27, Proposition 3.2]). *Let $\varepsilon_1, \varepsilon_2 \in (0,1/2)$ and $L \in \mathbb{N}$ be parameters. Then there are choices of (large) integers $K_1 = K_1(L,\varepsilon_1)$ and $K_2 = K_2(L,\varepsilon_1,\varepsilon_2)$ with the following property. For any set $S \subseteq \mathbb{Z}$, there is a decomposition of $S$ as a disjoint union $S_1 \amalg \ldots \amalg S_n \amalg S_0$ such that*

1. (Components are large) *$|S_i| \geqslant |S|/K_1$ for $i = 1, \ldots, n$;*
2. (Components are structured) *$\delta[S_i] \leqslant K_2$ for $i = 1, \ldots, n$;*
3. (Different components do not communicate) *$E(\lambda_i \cdot S_i, \lambda_j \cdot S_j) \leqslant \varepsilon_2|S_i|^{3/2}|S_j|^{3/2}$, whenever $1 \leqslant i < j \leqslant k$ and whenever $\lambda_i, \lambda_j \in \{1, \ldots, L\}$;*
4. (Noise term) *$E(\lambda_0 \cdot S_0, \lambda \cdot S) \leqslant \varepsilon_1|S|^3$ whenever $\lambda_0, \lambda \in \{1, \ldots, L\}$.*

*Observe that property (1) guarantees that $n \leqslant K_1$.*

This theorem is a structure theorem of sorts for sets of integers, making it similar in flavour to the arithmetic regularity lemma. However, they differ in the type structure they are seeking: this theorem exhibits additive structure in an arbitrary set, whereas the arithmetic regularity lemma is looking for analytic structure.

The quantity $T(S_1,S_2,S_3)$ is related to the additive energy via the following lemma.

**Lemma 3.2.3.** *Suppose that $S_1, S_2$ and $S_3 \subseteq \mathbb{Z}$ are finite sets. Then*

$$T(S_1,S_2,S_3)^2 \leqslant E(a_1 \cdot S_1, a_2 \cdot S_2)|S_3|.$$

*Proof.* For any $t \in \mathbb{Z}$, let $\mu(t)$ denote the number of ways of writing $t = a_1x_1 + a_2x_2$, for $x_i \in S_i$. Thus, by definition,

$$E(a_1 \cdot S_1, a_2 \cdot S_2) = \sum_t \mu(t)^2.$$

Now, we see that

$$T(S_1,S_2,S_3)^2 = \left(\sum_{t \in -a_3 \cdot S_3} \mu(t)\right)^2 \leqslant \left(\sum_t \mu(t)^2\right)|S_3| = E(a_1 \cdot S_1, a_2 \cdot S_2)|S_3|,$$

the inequality following from Cauchy-Schwarz. This completes the proof of Lemma 3.2.3. □

The following two facts are standard results in additive combinatorics.

**Lemma 3.2.4** (Ruzsa triangle inequality, [54, Lemma 2.6])**.** *For sets $A, B, C \subseteq \mathbb{Z}$,*

$$|A - C||B| \leqslant |A - B||B - C|.$$

**Lemma 3.2.5** (Energy Cauchy-Schwarz, [54, (2.9)])**.** *For sets $A, B \subseteq \mathbb{Z}$,*

$$E(A, B) \leqslant E(A, A)^{1/2} E(B, B)^{1/2}.$$

We will require the following lemma bounding $T(S_1, S_2, S_3)$.

**Lemma 3.2.6.** *Suppose that $S_1, S_2, S_3 \subseteq \mathbb{Z}$ are sets with sizes $s_1, s_2$ and $s_3$ respectively. Then, we have the bound*

$$T_{a_1,a_2,a_3}(S_1, S_2, S_3) \leqslant \frac{1}{4}\left(s_1 s_2 + s_2 s_3 + s_1 s_3 + 1\right). \tag{3.2.2}$$

*Proof.* We will first prove Lemma 3.2.6 in the case that $a_1, a_2$ and $a_3$ are all 1.

Without loss of generality, assume that $s_1 \leqslant s_2 \leqslant s_3$.

Suppose first that $s_3 \geqslant s_1 + s_2$. In that case, we have

$$\begin{aligned} T_{1,1,1}(S_1, S_2, S_3) &\leqslant s_1 s_2 \\ &\leqslant \frac{1}{4}(s_1 + s_2)^2 \\ &\leqslant \frac{1}{4}(s_1 s_3 + s_2 s_3). \end{aligned}$$

The first line follows from the trivial observation that for each pair of $x \in S_1$ and $y \in S_2$, there can be at most one solution to $x + y + z = 0$ with $z \in S_3$. The third line follows from our assumption on $s_3$. Thus, (3.2.2) follows in this case.

Now, suppose that $s_3 < s_1 + s_2$. In this case, we may apply [44, Lemma 2], which states that

$$T_{1,1,1}(S_1, S_2, S_3) \leqslant \frac{2(s_1 s_2 + s_2 s_3 + s_3 s_1) - (s_1^2 + s_2^2 + s_3^2) + 1}{4}. \tag{3.2.3}$$

(3.2.2) follows in this case via an easy application of the Cauchy-Schwarz inequality.

Finally, for arbitrary coefficients $a_1, a_2$ and $a_3$, observe that

$$\begin{aligned} T_{a_1,a_2,a_3}(S_1, S_2, S_3) &= T_{1,1,1}(a_1 \cdot S_1, a_2 \cdot S_2, a_3 \cdot S_3) \\ &\leqslant \frac{1}{4}\left(s_1 s_2 + s_2 s_3 + s_1 s_3 + 1\right). \end{aligned}$$

This completes the proof of Lemma 3.2.6. □

Finally, we will require the following theorem of Bukh:

**Theorem 3.2.7** ([9, Theorem 1.2])**.** *Given two coprime integers $\lambda_1$ and $\lambda_2$, we have that for any $S \subseteq \mathbb{Z}$,*

$$|\lambda_1 \cdot S + \lambda_2 \cdot S| \geqslant (|\lambda_1| + |\lambda_2|)|S| - o_{\lambda_1,\lambda_2}(|S|).$$

*Remark.* In fact, Shakan [52] has shown that we may take the $o_{\lambda_1,\lambda_2}(|S|)$ term to be constant. However, we will not gain anything by making use of this improvement, so will not do so.

### 3.2.2 Proof of Theorem 3.1.4

In this section, we will use the lemmas of Section 3.2.1 to prove Theorem 3.1.4. We must prove that, given a suitable choice of $a_1, a_2$ and $a_3$, all sufficiently large sets $S$ have $T(S) \leqslant \left(\frac{1}{12} + \varepsilon\right)|S|^2$.

Let $\varepsilon > 0$. Given our choice of $\varepsilon$, we must choose the values of the coefficients $a_1, a_2$ and $a_3$; we will do so later. Suppose that $S$ is a sufficiently large set. We will immediately apply the structure theorem, Theorem 3.2.2, to $S$, with $\varepsilon_1 = \left(\frac{\varepsilon}{6}\right)^4$, $\varepsilon_2 = 1$ and $L = 1$. This gives us a decomposition $S = S_1 \amalg \ldots \amalg S_n \amalg S_0$. We will start by showing that the contribution to $T(S)$ from solutions $a_1x_1 + a_2x_2 + a_3x_3 = 0$, with at least one of the $x_i$ taken from $S_0$, is small.

**Lemma 3.2.8.** *The number of solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$ in $S$, where some $x_i$ is taken from $S_0$, is no greater than $\frac{\varepsilon}{2}|S|^2$.*

*Proof.* The number of such solutions may be upper bounded by

$$T(S_0, S, S) + T(S, S_0, S) + T(S, S, S_0),$$

and so it suffices to show that each term is no greater than $\frac{\varepsilon}{6}|S|^2$.

Applying Lemmas 3.2.3 and 3.2.5, we have

$$\begin{aligned} T(S_0, S, S)^2 &\leqslant E(a_1 \cdot S_0, a_2 \cdot S)|S| \\ &\leqslant E(S_0, S_0)^{1/2}E(S, S)^{1/2}|S| \\ &\leqslant \varepsilon_1^{1/2}|S|^4, \end{aligned}$$

from which it follows that $T(S_0, S, S) \leqslant \varepsilon_1^{1/4}|S|^2$. By our choice of $\varepsilon_1$, this gives exactly what we claimed. □

At this point, we must bound the number of solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$ where each of $x_1, x_2$ and $x_3$ is taken from an $S_i$ with $i \geqslant 1$. To do this, we will start by restricting which triples $(i, j, k)$ can have the property that there are many solutions with $x_1 \in S_i, x_2 \in S_j$ and $x_3 \in S_k$. For instance, the fact that $\delta[S_1]$ is small, together with an assumption that $a_1$ and $a_2$ are coprime and $|a_1 + a_2|$ is large, will imply that there cannot be too many solutions with $x_1, x_2$ and $x_3$ all in $S_1$.

In particular, this will give us a fairly rigid structure on the collection of triples $S_i, S_j, S_k$ such that $T(S_i, S_j, S_k)$ can give a non-trivial contribution to $T(S, S, S)$. In order to quantify this structure, we will draw a labelled digraph $G$ whose vertices correspond to the $S_i$ with $i \geqslant 1$. We will draw an edge from $S_i$ to $S_j$ with label $\frac{a_1}{a_2}$ if and only if $T(S_i, S_j, S) \geqslant \frac{\varepsilon}{24K_1^2}|S|^2$, where $K_1$ is as in the statement of Theorem 3.2.2. Similarly, we will draw an edge with label $\frac{a_3}{a_2}$ if $T(S, S_j, S_i) \geqslant \frac{\varepsilon}{24K_1^2}|S|^2$, and similarly for the other four possible labels.

In particular, observe that if there is an edge from $S_i$ to $S_j$ with label $x$, then there will be an edge from $S_j$ to $S_i$ with label $x^{-1}$. Our definition of $G$ does not necessarily preclude the existence of multiple edges between $S_i$ and $S_j$ (with different labels), or edges from $S_i$ to $S_i$. However, as part of the proof, we will show that this cannot happen, provided that we assume a suitable hypothesis on $a_1, a_2$ and $a_3$.

First, we will show that $G$ captures almost all of the solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$.

**Lemma 3.2.9.** *Say that a triple $S_i, S_j$ and $S_k$ is* good *if and only if the six relevant edges are present. For example, $S_i \to S_j$ has label $\frac{a_1}{a_2}$, $S_k \to S_j$ has label $\frac{a_3}{a_2}$, and so on. Say that a triple is* bad *otherwise.*

*Then, the total number of solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$ among all of the bad triples is at most $\frac{\varepsilon}{4}|S|^2$.*

*Proof.* There are six ways a triple $(S_i, S_j, S_k)$ can be bad. One such way is if there is no edge from $S_i$ to $S_j$ with label $\frac{a_1}{a_2}$.

Let us count the total number of solutions among triples for which the $\frac{a_1}{a_2}$ edge is missing. That is

$$\sum_{\text{such } S_i, S_j} T(S_i, S_j, S) \leqslant K_1^2 \frac{\varepsilon}{24K_1^2}|S|^2$$
$$= \frac{\varepsilon}{24}|S|^2,$$

since the number of pairs $S_i, S_j$ is bounded by $K_1^2$.

Summing this over the six possible ways for a triple to be bad completes the proof of Lemma 3.2.9. □

In view of Lemmas 3.2.8 and 3.2.9, it remains to show that the number of solutions among the good triples is at most $\left(\frac{1}{12}+\frac{\varepsilon}{4}\right)|S|^2$, for a suitable choice of the coefficients $a_1, a_2$ and $a_3$. The values we will choose are $a_1 = 1, a_2 = M$ and $a_3 = M+1$, where $M$ is any positive integer satisfying

$$M > \left(\frac{1000K_1^4K_2^2}{\varepsilon^2}\right)^{K_1}. \tag{3.2.4}$$

We can now prove the following lemma:

**Lemma 3.2.10.** *With the values of $a_1, a_2$ and $a_3$ that we have chosen, the product of the labels along any cycle in $G$ must be 1.*

*Remark.* This immediately tells us that $G$ has no loops (edges from a vertex to itself). In view of the fact that an edge from $S_i$ to $S_j$ with label $x$ is accompanied by an edge from $S_j$ to $S_i$ with label $x^{-1}$, this also tells us that there can be at most one edge from $S_i$ to $S_j$.

*Remark.* We have chosen particular values of the $a_i$ for simplicity; indeed, we only need a single choice of coefficients to work in order to establish Theorem 3.1.4. However, the same argument is able to establish Lemma 3.2.10, and thus also Theorem 3.1.4, for a much wider class of equations. For example, whenever $a_1, a_2$ and $a_3$ are coprime, and at least two of the three coefficients are large enough, then the analogue of Lemma 3.2.10 holds, and thus $\gamma_{a_1,a_2,a_3} < 1/12 + \varepsilon$.

Conversely, it does not suffice for just one of the $a_i$ to be large. For example, if $a_1 = a_2 = 1$, then it can be shown that, for $S$ a slightly modified version of the set in Proposition 3.1.2, $T_{1,1,a_3}(S) > \frac{1}{5}|S|^2$ for any $a_3$; in particular, we can take $S = a_3[-3M, 3M] \cup [-2M, 2M]$. We will discuss this further in Section 3.9.

The idea behind Lemma 3.2.10 is as follows. Suppose that, instead of the condition that the $S_i$ have small doubling, we had the stronger condition that each $S_i$ was an arithmetic progression. Then the assertion that there is an edge from $S_i$ to $S_j$ with label $t$ implies that the ratio between the common differences of $S_i$ and $S_j$ is close to $t^{-1}$. In this setting, it is obvious that the product of the edge labels around a cycle must be 1; indeed, the product of the edge labels along a path is the ratio between the common differences of the endpoints.

Of course, it is not necessarily the case that each $S_i$ is an arithmetic progression. However, our proof will use this idea, by following the product of the labels along a path and relating it to the endpoints of the path, in such a way that if the path is a cycle, that product must be 1.

*Proof of Lemma 3.2.10.* Suppose there is a cycle whose label product is not 1; consider a shortest such cycle. By minimality, such a cycle may have no repeated vertices, and thus must have at most $K_1$ vertices. Thus, without loss of generality the cycle is $S_1, S_2, \ldots, S_k, S_1$, where $S_i \to S_{i+1}$ has label $t_i$ (with $S_{k+1} = S_1$), and $k \leqslant K_1$.

By Lemma 3.2.3, we deduce that for each $i$,

$$E(t_i \cdot S_i, S_{i+1}) \geqslant \frac{\varepsilon^2}{576K_1^4}|S|^3.$$

Now, let us apply Lemma 3.2.1 to $S_i$ and $S_{i+1}$. We have that

$$E(t_i \cdot S_i, S_{i+1}) \geqslant \frac{\varepsilon^2}{576K_1^4}|S_i|^{3/2}|S_{i+1}|^{3/2},$$

and so we deduce that

$$\begin{aligned} |t_i \cdot S_i - S_{i+1}| &\leqslant \delta[S_i]\delta[S_{i+1}]|S_i|^{1/2}|S_{i+1}|^{1/2}\frac{576K_1^4}{\varepsilon^2} \\ &\leqslant \frac{576K_1^4K_2^2}{\varepsilon^2}|S_i|^{1/2}|S_{i+1}|^{1/2}. \end{aligned} \tag{3.2.5}$$

Now, we can prove, by inductively applying Lemma 3.2.4, that

$$|t_1t_2\ldots t_i \cdot S_1 - S_{i+1}| \leqslant \left(\frac{576K_1^4K_2^2}{\varepsilon^2}\right)^i |S_1|^{1/2}|S_{i+1}|^{1/2}. \tag{3.2.6}$$

Thus, setting $i = k$, we learn that

$$|t_1t_2\ldots t_k \cdot S_1 - S_1| \leqslant \left(\frac{576K_1^4K_2^2}{\varepsilon^2}\right)^{K_1} |S_1|, \tag{3.2.7}$$

since $k \leqslant K_1$.

By hypothesis, $t_1t_2\ldots t_k \neq 1$. However, we know that $t_1t_2\ldots t_k$ can be written in the form $M^{e_1}(M+1)^{e_2}$ for some integers $e_i$ not both zero. Suppose that $e_1$ is nonzero; the argument is similar if $e_2$ is nonzero.

Write $t_1t_2\ldots t_k = \frac{r}{s}$ for coprime integers $r$ and $s$; our hypothesis tells us that $M$ must divide $r$ or $s$. Therefore,

$$|r| + |s| > \left(\frac{1000K_1^4K_2^2}{\varepsilon^2}\right)^{K_1},$$

as a consequence of (3.2.4).

Thus, we have shown that $|r \cdot S_1 - s \cdot S_1| \leqslant \left(\frac{576K_1^4K_2^2}{\varepsilon^2}\right)^{K_1}|S_1|$. But, if $S_1$ is sufficiently large, this contradicts Theorem 3.2.7, which states that

$$|r \cdot S_1 - s \cdot S_1| > \left(\frac{1000K_1^4K_2^2}{\varepsilon^2}\right)^{K_1} |S_1|,$$

whenever $|S_1|$ is sufficiently large.

This contradiction completes the proof of Lemma 3.2.10. □

To complete the proof of Theorem 3.1.4, we just need to bound the number of solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$, with $x_1, x_2, x_3$ taken from a good triple. The following lemma will achieve this.

**Lemma 3.2.11.** *Suppose we choose $a_1 = 1, a_2 = M$ and $a_3 = M + 1$, as in the statement of Lemma 3.2.10.*

*Then the number of solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$ taken from good triples is bounded above by $\left(\frac{1}{12} + \frac{\varepsilon}{4}\right)|S|^2$, whenever $|S|$ is large enough.*

*Proof.* We will start by defining a function

$$d : [n] \to \mathbb{Q}, \tag{3.2.8}$$

with the property that if $S_i \to S_j$ has label $t$, then $d(j) = td(i)$.

One way we can do this is as follows. For each connected component $G'$ of $G$, choose the smallest value of $i$ such that $S_i$ is in $G'$, and set $d(i) = 1$. Then, for any other $j$ with $S_j$ in $G'$, $d(j)$ is determined by the product of the labels on any path from $S_i$ to $S_j$. Lemma 3.2.10 guarantees that this value does not depend on the path chosen.

Now, for each $d$, let $R_d = \cup_{i:d(i)=d} S_i$. Suppose that $S_i, S_j, S_k$ is a good triple, in that order (so, for example, the label on $S_i \to S_j$ is $a_1/a_2$). Then, setting $d = a_1d(i)$, we have that $d(i) = d/a_1, d(j) = d/a_2$ and $d(k) = d/a_3$.

Therefore, all of the solutions coming from the good triple $S_i, S_j, S_k$ will be counted in $T(R_{d/a_1}, R_{d/a_2}, R_{d/a_3})$, and so an upper bound for the total number of solutions coming from good triples is

$$\sum_d T(R_{d/a_1}, R_{d/a_2}, R_{d/a_3}),$$

where the sum is taken over all $d$ such that all three of the $R_i$ exist (in particular, there can be no more than $n$ terms in the sum).

We may apply Lemma 3.2.6 to give an upper bound for this.

$$\begin{aligned}\sum_d T(R_{d/a_1}, R_{d/a_2}, R_{d/a_3}) &\leqslant \frac{1}{4}\sum_d |R_{d/a_1}||R_{d/a_2}| + |R_{d/a_1}||R_{d/a_3}| + |R_{d/a_2}||R_{d/a_3}| + 1 \\ &\leqslant \frac{1}{4}\sum_{d_1 \sim d_2} |R_{d_1}||R_{d_2}| + K_1, \end{aligned} \tag{3.2.9}$$

where the sum on the second line is over *unordered* pairs $d_1, d_2$ such that $d_1/d_2$ is equal to the ratio between two of the $a_i$. The second inequality follows because if

$d_1 \sim d_2$, then there is exactly one ratio $a_i/a_j$ such that $d_1/d_2 = a_i/a_j$. Thus, the term $|R_{d_1}||R_{d_2}|$ appears in at most one of the sums on the right hand side of the first line.

Finally, for $i = 0, 1$ and 2, define the quantity $X_i$ by

$$X_i = \sum_{\substack{d=M^{e_1}(M+1)^{e_2} \\ e_1-e_2=i \mod 3}} |R_d|.$$

By our construction of $d$, each $|R_d|$ appears as a term in exactly one of the $X_i$. Furthermore, $d_1 \sim d_2$ only if $R_{d_1}$ and $R_{d_2}$ are in different sums $X_i$, and any term $|R_{d_1}||R_{d_2}|$ appears at most once in (3.2.9). Consequently, we have the upper bound

$$\begin{aligned}\frac{1}{4}\sum_{d_1\sim d_2}|R_{d_1}||R_{d_2}| &\leqslant \frac{1}{4}(X_0X_1 + X_0X_2 + X_1X_2) \\ &\leqslant \left(\frac{1}{12} + \frac{\varepsilon}{4}\right)|S|^2,\end{aligned}$$

the latter inequality following from an easy application of Cauchy-Schwarz, since $X_0 + X_1 + X_2 \leqslant |S|$. This completes the proof of Lemma 3.2.11. □

*Remark.* Recall that the construction for a set $S$ with $T(S) \geqslant \frac{1}{12}|S|^2 + O(|S|)$ given in Proposition 3.1.2 consisted of three arithmetic progressions $S_1, S_2$ and $S_3$, with common differences in the ratio $1/a_1 : 1/a_2 : 1/a_3$. In this instance, the vertices of $G$ would be $S_1, S_2$ and $S_3$ and there would be all six edges between the $S_i$, with the appropriate labels. In particular, we would have $d(1) = 1$, $d(2) = a_1/a_2$ and $d(3) = a_1/a_3$; in other words, the values of $d(i)$ are proportional to the common differences of the arithmetic progressions that are the $S_i$.

We have now essentially proven Theorem 3.1.4. Indeed, any solution to $a_1x_1 + a_2x_2 + a_3x_3 = 0$ must either have some $x_i$ in $S_0$, or must come from a bad triple, or must come from a good triple. Combining Lemmas 3.2.8, 3.2.9 and 3.2.11 gives the result if $|S|$ is large enough. □

## 3.3 Proving Lemma 3.1.5 via arithmetic regularity

In this section, we will prove Lemma 3.1.5. The proof involves three key ingredients. We will first need to exhibit some degree of structure in a set $S$ with $T(S)$ large. Indeed, we will show that such an $S$ can essentially be taken to be a dense subset of an interval $[-M/2, M/2]$, in Section 3.3.1.

Once we have done this, we will be able to apply the arithmetic regularity lemma, introduced in [25]. The particular version we will need is the $s = 1$ case of [29,

Theorem 1.2]; a convenient exposition of this is given in [16]. With this, we will be able to show that, for any set $S$ of integers, calculating $T(S)/|S|^2$ is essentially equivalent to a similar calculation for a function over $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$, where $\mathbb{T}$ denotes $\mathbb{R}/\mathbb{Z}$. The details of the application of the arithmetic regularity lemma will form Section 3.3.2.

After we have reduced to the situation of $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$, we will exploit various concavity properties to show the following. Given our function $f$ defined on $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$, for any fixed $x \in \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$ we can take $f(x, \cdot) : [-1/2, 1/2] \to [0, 1]$ to be the indicator function of an interval centred at 0. Once we have done this, we will be able to show that the length of this interval can be taken to be independent of the $\mathbb{T}^d$ component. This will essentially complete the proof of Lemma 3.1.5, and will form Section 3.3.3.

### 3.3.1 Finding a structured subset

We will now begin the proof of Lemma 3.1.5. We wish to prove that, in the calculation of $\gamma = \gamma_{a_1,a_2,a_3}$, it is enough to consider dense subsets of $[-M/2, M/2]$. In particular, we will establish the following proposition.

**Proposition 3.3.1.** *Fix any $\varepsilon > 0$. Then, there exists a constant $\eta > 0$ with the following property. We can find arbitrarily large sets $S'$ with $t(S') \geqslant \gamma - \varepsilon$, and such that for some $M \in \mathbb{N}$, $S' \subseteq [-M/2, M/2]$ and $|S'| \geqslant \eta M$.*

It will be necessary to establish this proposition before we can apply the arithmetic regularity lemma.

The proof of Proposition 3.3.1 will fall into two steps. Firstly, given a large set $S$ with $t(S)$ large, say $t(S) \geqslant \gamma - \frac{\varepsilon}{2}$, we will find a large subset $S_1 \subseteq S$ with $t(S_1) \geqslant \gamma - \varepsilon$, and for which $S_1$ has a high degree of additive structure. We will use Theorem 3.2.2 to accomplish this. Once we have done this, we will be able to apply Freiman's Theorem to $S_1$. This will give a bijection between $S_1$ and a subset $S'$ of $[-M, M]$, which preserves solutions to (3.1.1).

We will begin the proof of Proposition 3.3.1 with a lemma.

**Lemma 3.3.2.** *Choose $\varepsilon > 0$. Then there is a constant $D \ll_\varepsilon 1$ with the following property. Suppose that $S$ is a large set with $t(S) \geqslant \gamma - \frac{\varepsilon}{2}$. Then $S$ has a subset $S_1$,*

*with*

$$|S_1| \geqslant |S|/D,$$
$$\delta[S_1] \leqslant D,$$
$$t(S_1) \geqslant \gamma - \varepsilon.$$

*Remark.* Observe that we can take $|S|$ to be arbitrarily large. In other words, we immediately obtain the following corollary.

**Corollary** (Corollary to Lemma 3.3.2)**.** *Choose $\varepsilon > 0$. Then for the same constant $D \ll_\varepsilon 1$, there must be arbitrarily large sets $S$, with $\delta[S] \leqslant D$ and $t(S) \geqslant \gamma - \varepsilon$.*

*Proof of Lemma 3.3.2.* We will start by applying Theorem 3.2.2 to $S$. Choose $L = |a_1| + |a_2| + |a_3|$, and let $\varepsilon_1$ and $\varepsilon_2$ be small positive constants to be specified later. Applying Theorem 3.2.2 to $S$ gives us a decomposition

$$S = S_1 \cup \cdots \cup S_n \cup S_0.$$

It will be sufficient to prove that, for at least one of the values of $i \neq 0$, $t(S_i) \geqslant \gamma - \frac{\varepsilon}{2}$. This will establish Lemma 3.3.2, with $D > F_1(L, \varepsilon_1)$ and $D > F_2(L, \varepsilon_1, \varepsilon_2)$. Given our decomposition $S = S_1 \cup \cdots \cup S_n \cup S_0$, observe that

$$T(S) = T(S, S, S) = \sum_{i_1, i_2, i_3} T(S_{i_1}, S_{i_2}, S_{i_3})$$

over all triples $i_1, i_2, i_3$. We will show that the main contribution to this sum is from $k$-tuples with $i_1 = i_2 = i_3 \neq 0$.

*Case 1: $i_j = 0$ for some $j$.* Let us start by bounding the contribution from terms for which $i_1 = 0$. The sum of such terms is bounded by $T(S_0, S, S)$. We can use Lemma 3.2.3 to bound these terms:

$$T(S_0, S, S)^2 \leqslant E(a_1 \cdot S_0, a_2 \cdot S)|S| \leqslant \varepsilon_1 \times |S|^4,$$

where the second inequality follows from property 4 of Theorem 3.2.2. Thus, we have $T(S_0, S, S) \leqslant \varepsilon_1^{1/2}|S|^2$. Summing this up over each choice of $j$ with $i_j = 0$ , we see that the total contribution from such terms is bounded by $3\varepsilon_1^{1/2}|S|^2$.

*Case 2: the $i_j$ are nonzero and not all equal.* For terms with $i_1 \neq i_2$, again we use Lemma 3.2.3 to see that

$$T(S_{i_1}, S_{i_2}, S_{i_3})^2 \leqslant E(a_1 \cdot S_{i_1}, a_2 \cdot S_{i_2})|S_{i_3}|.$$

Therefore, property 3 of Theorem 3.2.2 tells us that

$$T(S_{i_1}, S_{i_2}, S_{i_3}) \leqslant (\varepsilon_2)^{1/2}|S_{i_1}|^{3/4}|S_{i_2}|^{3/4}|S_{i_3}|^{1/2}.$$

By Hölder's inequality, it follows that the sum of all such terms (with $i_1 \neq i_2$) is bounded above by $n\varepsilon_2^{1/2}|S|^2$. Thus, the total contribution from terms where the $i_j$ are not all equal is at most $3n\varepsilon_2^{1/2}|S|^2$.

Thus, the total contribution from terms $T(S_{i_1}, S_{i_2}, S_{i_3})$, other than those for which $i_1 = i_2 = i_3 \neq 0$, is at most $(3\varepsilon_1^{1/2} + 3n\varepsilon_2^{1/2})|S|^2$. Since $n$ is bounded independently of $\varepsilon_2$ in Theorem 3.2.2, this will certainly be less than $\frac{\varepsilon}{2}|S|^2$, provided that $\varepsilon_1$ and $\varepsilon_2$ are small enough.

Now, suppose that for each of the $S_i$, $t(S_i, S_i, S_i) < \gamma - \varepsilon$, or, in other words, $T(S_i, S_i, S_i) < (\gamma - \varepsilon)|S_i|^2$. Then,

$$\begin{aligned} T(S) &= \sum_{i_1,i_2,i_3} T(S_{i_1}, S_{i_2}, S_{i_3}) \\ &\leqslant \sum_i T(S_i, S_i, S_i) + \frac{\varepsilon}{2}|S|^2 \\ &< (\gamma - \varepsilon)\sum_i |S_i|^2 + \frac{\varepsilon}{2}|S|^2 \\ &\leqslant \left(\gamma - \frac{\varepsilon}{2}\right)|S|^2 \end{aligned}$$

which contradicts our assumption on the size of $t(S)$.

Thus, there must be a component $S_i$ with $t(S_i) \geqslant \gamma - \varepsilon$, and, without loss of generality, we may assume that that component is $S_1$. □

*Proof of Proposition 3.3.1.* Thanks to the corollary following Lemma 3.3.2, we can find arbitrarily large sets $S$ with

$$\begin{aligned} \delta[S] &\leqslant D \\ t(S) &\geqslant \gamma - \varepsilon \end{aligned}$$

for some $D$ depending only on $a_1, a_2, a_3,$ and $\varepsilon$.

Recall that $L = |a_1| + |a_2| + |a_3|$. A Freiman $L$-isomorphism is an injective map $f$ for which $a_1 + \cdots + a_L = b_1 + \cdots + b_L$ holds if and only if $f(a_1) + \cdots + f(a_L) = f(b_1) + \cdots + f(b_L)$ holds. In particular, if $a_1x_1 + a_2x_2 + a_3x_3 = 0 = a_1 \cdot 0 + a_2 \cdot 0 + a_3 \cdot 0$, then $a_1f(x_1) + a_2f(x_2) + a_3f(x_3) = (a_1 + a_2 + a_3)f(0)$. Replacing $f(x)$ with $f(x) - f(0)$ if necessary, we see that $f$ is a map that preserves solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$. Therefore, it will suffice to find a Freiman $L$-isomorphism from $S \cup \{0\}$ to a subset of $[-M/2, M/2]$, for some $M$ not much larger than $|S|$.

To do this, let us begin by applying Freiman's theorem [48, Theorem 1.1] to $S$. Because $\delta[S] \ll_\varepsilon 1$ (recalling that $\delta[S] = |S - S|/|S|$), we obtain a degree $d \ll_\varepsilon 1$ generalised arithmetic progression $P$ of size $|P| \ll_\varepsilon |S|$ containing $S$. More precisely, we obtain integers $a$, $b_1, \dots, b_d$ and $h_1, \dots, h_d$ such that

$$P = a + h_1[-b_1, b_1] + \cdots + h_d[-b_d, b_d]$$

contains $S$.

Next, apply [54, Theorem 3.40] to $P$. This gives us a generalised arithmetic progression $P'$ containing $P$, with $|P'| \ll |P|$ and $\dim P' = d' \leqslant d$, which is *$L$-proper*. Informally, this means that any triple $x_1, x_2, x_3$ of elements of $P'$ with $a_1x_1 + a_2x_2 + a_3x_3 = 0$ must have $a_1x_1 + a_2x_2 + a_3x_3 = 0$ *in* $P'$. More formally, suppose we write

$$P' = a' + h'_1[-b'_1, b'_1] + \cdots + h'_{d'}[-b'_{d'}, b'_{d'}],$$

and have a solution $x_1, x_2, x_3$ with $x_i = a' + h'_1x_{i1} + \cdots + h'_{d'}x_{id'}$. Then, for each $1 \leqslant j \leqslant d'$, $a_1x_{1j} + a_2x_{2j} + a_3x_{3j} = 0$. Without any loss of generality, we may assume that $P$ itself is $L$-proper.

Finally, we may complete the proof by constructing the natural $L$-isomorphism

$$a + h_1y_1 + \cdots + h_dy_d \to y_1 + B_2y_2 + B_3y_3 + \cdots + B_dy_d,$$

where $B_i = 2^i b_1b_2 \dots b_{i-1}$.

The image of this map is contained within an interval of length $2^{d+1}b_1b_2 \dots b_d \ll_d |P| \ll_\varepsilon |S|$, where $d \ll_\varepsilon 1$. In other words, we have exactly a Freiman $L$-isomorphism from $S$ to a subset $S'$ of $[-M/2, M/2]$, where $|S'| \gg_\varepsilon M$. □

### 3.3.2 Applying the arithmetic regularity lemma

In this section, we will use the arithmetic regularity lemma [29, Theorem 1.2] to reduce the situation from subsets of $\mathbb{Z}$ to functions defined on $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$. For any $\varepsilon > 0$, Proposition 3.3.1 guarantees us a constant $\eta > 0$ such that we can find arbitrarily large values of $M$ and sets $S$, such that $S \subseteq [-M/2, M/2]$ has density at least $\eta$, and such that $t(S) \geqslant \gamma - \varepsilon$. The arithmetic regularity lemma provides a way to classify the structure of a dense subset $S$ of an interval, which will make it easier for us to find bounds on $t(S)$. It will be most convenient for us in the following form:

**Theorem 3.3.3** (Arithmetic Regularity Lemma)**.** *Let $f : [-M/2, M/2] \to [0, 1]$ be a function, $\mathcal{F}$ an increasing function, and $\delta > 0$. Then there is a quantity $B \ll_{\delta,\mathcal{F}} 1$ and a decomposition*

$$f = f_{str} + f_{sml} + f_{unf}$$

*of $f$ into functions $f_{str}, f_{sml}, f_{unf} : [-M/2, M/2] \to [-1, 1]$ such that*

1. *$f_{str}(n) = F(n \mod q, n/M, \theta n)$, where*

$$F : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d \to [0, 1],$$

   *$q, d \leqslant B$, $F$ has Lipschitz constant at most $B$, and $\theta \in \mathbb{T}^d$ is $(\mathcal{F}(B), M)$-irrational (see below),*

2. *$\sum_{-M/2 \leqslant n \leqslant M/2} |f_{sml}(n)| \leqslant \delta M$,*

3. *$\sup_t |\hat{f}_{unf}(t)| \leqslant M/\mathcal{F}(B)$,*

4. *$f_{str}$ and $f_{str} + f_{sml}$ take values in $[0, 1]$.*

*Here, $\hat{f}(t) = \sum_{-M/2 \leqslant n \leqslant M/2} f(n) \exp(-2\pi i t n)$ is the Fourier Transform of $f$ on $\mathbb{Z}$. The statement that $\theta$ is $(\mathcal{F}(B), M)$-irrational means that, for any $r \in \mathbb{Z}^d$ with $\sum |r_i| \leqslant \mathcal{F}(B)$ satisfies $\sum ||r_i \theta_i|| \geqslant \mathcal{F}(B)/M$, where $||x||$ is the distance from $x$ to the nearest integer.*

As in [26], we can control $f_{\text{sml}}$ with the $\ell^1$ norm instead of the $\ell^2$ norm as $|f_{\text{sml}}|$ is bounded by 1; similarly, we can control $f_{\text{unf}}$ with the supremum norm of the Fourier transform instead of the Gowers $U^2$ norm.

Fix $\varepsilon > 0$ small. Then, we will consider large values of $M$ and subsets $S \subseteq [-M/2, M/2]$ with density $\geqslant \eta$ and $t(S) \geqslant \gamma - \varepsilon$; such subsets are guaranteed to exist by Proposition 3.3.1.

Let $\delta > 0$ be a constant and $\mathcal{F}$ a growth function to be specified later, and let $f$ be the indicator function of $S$. We will immediately apply the arithmetic regularity lemma to $f$. This gives us a decomposition

$$f = f_{\text{str}} + f_{\text{sml}} + f_{\text{unf}}$$

as described in the statement of Theorem 3.3.3.

We will first show that the contributions from the error terms $f_{\text{sml}}$ and $f_{\text{unf}}$ may be ignored.

**Claim.** *For any $\varepsilon' > 0$, suppose that $\delta$ is small enough, and $\mathcal{F}$ grows quickly enough (both depending on $\varepsilon'$), in the statement of Theorem 3.3.3; it will suffice to have $\delta < \varepsilon'/100$ and $\mathcal{F}(B) > 100/\varepsilon'$. Then, we have the following two bounds:*

$$\left| \sum_{-M/2 \leqslant n \leqslant M/2} f_{str}(n) - \sum_{-M/2 \leqslant n \leqslant M/2} f(n) \right| \leqslant \varepsilon' M \tag{3.3.1}$$

$$|T(f_{str}) - T(f)| \leqslant \varepsilon' M^2. \tag{3.3.2}$$

*Proof of Claim.* Establishing (3.3.1) is easy; we have that $\sum |f_{\mathrm{sml}}| \leqslant \delta M$, and that $\sum f_{\mathrm{unf}} = \hat{f}_{\mathrm{unf}}(0) \leqslant M/\mathcal{F}(B)$.

For (3.3.2), we start by writing $f = f_{\mathrm{str}} + f_{\mathrm{sml}} + f_{\mathrm{unf}}$, and we may expand out $T(f) = T(f,f,f)$ as a sum of 27 terms. One of these terms is $T(f_{\mathrm{str}}, f_{\mathrm{str}}, f_{\mathrm{str}})$. All of the other terms have three arguments to $T$, at least one of which is $f_{\mathrm{sml}}$ or $f_{\mathrm{unf}}$.

**Terms where one of the arguments is $f_{\mathrm{sml}}$.** Without loss of generality, suppose that the first argument is $f_{\mathrm{sml}}$, and write $f_2$ and $f_3$ for the other arguments. Then, by definition of $T$,

$$\begin{aligned} T(f_{\mathrm{sml}}, f_2, f_3) &\leqslant \sum_{n_1,n_2} |f_{\mathrm{sml}}(n_1) f_2(n_2)| \\ &\leqslant \left(\sum_{n_1} |f_{\mathrm{sml}}(n_1)|\right)\left(\sum_{n_2} |f_2(n_2)|\right) \\ &\leqslant \delta M^2 \end{aligned}$$

as required.

**Terms where one of the arguments is $f_{\mathrm{unf}}$.** As before, suppose that the first argument is $f_{\mathrm{unf}}$, and write $f_2$ and $f_3$ for the other arguments.

Now, we can write

$$T(f_1,f_2,f_3) = \sum_{\substack{|n_1|,|n_2|,|n_3|\leqslant M \\ a_1n_1+a_2n_2+a_3n_3=0}} f_1(n_1)f_2(n_2)f_3(n_3).$$

For each $i$, write $f_i'(a_i x) = f_i(x)$ (and set $f_i'(x) = 0$ if $a_i \nmid x$). Then $f_1'$ still satisfies property 3 of Theorem 3.3.3 (in the sense that $\sup_t \hat{f}_1'(t) \leqslant a_1 M/\mathcal{F}(B)$). And, we now have

$$\begin{aligned} T(f_1,f_2,f_3) &= \sum_{\substack{|n_1|,|n_2|,|n_3|\leqslant LM \\ n_1+n_2+n_3=0}} f_1'(n_1)f_2'(n_2)f_3'(n_3) \\ &= \int_0^1 \hat{f}_1'(t)\hat{f}_2'(t)\hat{f}_3'(t)dt \\ &\leqslant \sup_t \left|\hat{f}_1'(t)\right| \left(\int_0^1 |\hat{f}_2'(t)|^2 dt\right)^{1/2} \left(\int_0^1 |\hat{f}_3'(t)|^2 dt\right)^{1/2}, \end{aligned}$$

the last line following from Cauchy-Schwarz. The first factor is bounded by $M/\mathcal{F}(B)$ as $f_1 = f_{\mathrm{unf}}$, and the second and third can be bounded by $M^{1/2}$ thanks to Parseval's identity. Thus, $T(f_{\mathrm{unf}}, f_2, f_3) \ll M^2/\mathcal{F}(B)$. This completes the proof of the claim. $\Diamond$

Thus, we have a function $f_{\mathrm{str}}$ which can be written in the form described by property 1 of Theorem 3.3.3. As is standard in applications of the Regularity Lemma, we will use a counting lemma to prove a relation of the form $t(f_{\mathrm{str}}) = (1+o(1))t(F)$. Indeed, [17, Lemma A.4] states that provided that $\mathcal{F}(B)$ and $M$ are large enough, then

$$\left|\frac{1}{M}\sum_{n} f_{\mathrm{str}}(n) - \int F\right| \leqslant \varepsilon', \tag{3.3.3}$$

where the integral $\int F$ refers to the integral over $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$ with the uniform measure, normalised so that $\int 1 = 1$.

We would like to define a version of $T$ on $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$ in such a way that a relation similar to (3.3.3) holds for $T$; in other words, that $T(f_{\mathrm{str}}) \approx T(F)$. The current expression as a sum over all solutions to (3.1.1) is fairly unwieldy in the local setting. One way to do this is to parametrise the set of solutions to the equation with a two parameter family. We do this with the following proposition.

**Proposition 3.3.4** (Parametrisation of solutions)**.** *Regardless of the values of the coefficients $(a_1, a_2, a_3)$ (which we assume to be coprime and nonzero), there exists a pair of integer solutions $(y_1, y_2, y_3)$ and $(z_1, z_2, z_3)$ to (3.1.1) such that, for any integer solution $(x_1, x_2, x_3)$ to (3.1.1), there are integers $u$ and $v$ such that $x_i = y_i u + z_i v$ for each $i$.*

*Furthermore, we have the following two properties.*

- *Regardless of the choice of basis,*

$$\det \begin{pmatrix} y_1 & z_1 \\ y_2 & z_2 \end{pmatrix} = \pm a_3. \tag{3.3.4}$$

- *Let $G$ be any abelian group, and let $x_1, x_2$ and $x_3$ be any solution to (3.1.1) in $G$. Then, there exist $u, v \in G$ so that $x_i = y_i u + z_i v$ for each $i$, where $y_i$ and $z_i$ form our solution basis over $\mathbb{Z}$ from Proposition 3.3.4.*

*Proof.* The existence of the solutions $(y_i)$ and $(z_i)$ such that any solution is an integer linear combination of them follows from the fact that any torsion-free finite rank abelian group has a basis. Indeed, the set of solutions to (3.1.1) is a rank two subgroup of $\mathbb{Z}^3$, and we may take $(y_i)$ and $(z_i)$ to be a basis.

To establish (3.3.4), we observe that the set of solutions $x_1, x_2, x_3$ to (3.1.1) all satisfy the relation

$$\det \begin{pmatrix} y_1 & z_1 & x_1 \\ y_2 & z_2 & x_2 \\ y_3 & z_3 & x_3 \end{pmatrix} = 0,$$

as the last column is a linear combination of the first two. We may expand out the determinant to obtain the relation

$$\det \begin{pmatrix} y_2 & z_2 \\ y_3 & z_3 \end{pmatrix} x_1 + \det \begin{pmatrix} y_3 & z_3 \\ y_1 & z_1 \end{pmatrix} x_2 + \det \begin{pmatrix} y_1 & z_1 \\ y_2 & z_2 \end{pmatrix} x_3 = 0,$$

so the coefficients of this relation must match $(a_1, a_2, a_3)$ up to scaling by a constant. However, these coefficients must be coprime, because the elements $(y_1, y_2, y_3)$ and $(z_1, z_2, z_3)$ of $\mathbb{Z}^3$ may be extended to a basis by any $(w_1, w_2, w_3)$ satisfying the relation $\sum_i a_i w_i = 1$ (such $(w_i)$ must exist as the coefficients $a_i$ are coprime). Thus, the coefficient of $x_3$ can only be equal to $a_3$ or $-a_3$, as required.

To prove the final part, we will need to have integers $(w_i)$ such that the matrix

$$M = \begin{pmatrix} y_1 & z_1 & w_1 \\ y_2 & z_2 & w_2 \\ y_3 & z_3 & w_3 \end{pmatrix}$$

satisfies $\det M = 1$. To do this, we use the $(w_i)$ from before, possibly flipping the sign of each of them if necessary. In particular, since the determinant of $M$ is 1, it must have an inverse $M^{-1}$ over $\mathbb{Z}$.

Now, given a solution $x_1, x_2, x_3$ to (3.1.1) in an abelian group $G$, define $u, v$ and $\alpha$ by

$$\begin{pmatrix} u \\ v \\ \alpha \end{pmatrix} = M^{-1} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The formula for the inverse of a $3 \times 3$ matrix, combined with (3.3.4) and its cyclic shifts, tells us that the bottom row of $M^{-1}$ is $\pm(a_1, a_2, a_3)$; in particular, $\alpha = 0$. Consequently, the $u$ and $v$ produced by this procedure satisfy the required relation $x_i = uy_i + vz_i$. □

Now that we have a parametrisation for the set of solutions to (3.1.1), we may rewrite the definition of $T(f)$ for $f : \mathbb{Z} \to [0, 1]$ a compactly supported function. Indeed, we have

$$T(f) = \sum_{u,v \in \mathbb{Z}} f(y_1 u + z_1 v) f(y_2 u + z_2 v) f(y_3 u + z_3 v)., \tag{3.3.5}$$

as the right hand side is exactly the same as the sum of $f(x_1)f(x_2)f(x_3)$ over triples $x_1, x_2, x_3$ satisfying (3.1.1). Furthermore, it is now possible to sensibly define $T(F)$ for functions $F : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d \to [0, 1]$. Indeed, we extend $F$ to the abelian

group $\mathbb{Z}/q\mathbb{Z} \times \mathbb{R} \times \mathbb{Z}^d$ such that $F(x) = 0$ if the component of $x$ in $\mathbb{R}$ is not within the interval $[-1/2, 1/2]$. This allows us to use the same formula above for $T$:

$$T(F) = \iint F(y_1u + z_1v)F(y_2u + z_2v)F(y_3u + z_3v)dudv, \tag{3.3.6}$$

where the integrals are normalised as before (in other words, so that the integral over $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$ is 1).

Having defined $T$ this way, we now have the bound

$$\left|\frac{1}{M^2}T(f_{\mathrm{str}}) - T(F)\right| \leqslant \varepsilon', \tag{3.3.7}$$

provided that $\mathcal{F}(B)$ and $M$ are large enough. This is a counting lemma for $T$; its proof is similar to the proof of [17, Lemma A.5]; indeed, in that paper it is shown that if $\pi : \mathbb{Z} \to \mathbb{Z}/q\mathbb{Z} \times \mathbb{R} \times \mathbb{T}^d$ is defined by $\pi(n) = (n \mod q, n/M, \theta n)$, then we have that for any Lipschitz $F_*$,

$$\frac{1}{M^2}\sum_{u,v\in[-M/2,M/2]} F_*(\pi(u), \pi(v)) \approx \iint F_*(u, v)dudv.$$

Applying this to $F_*(u, v) = F(y_1u + z_1v)F(y_2u + z_2v)F(y_3u + z_3v)$ gives (3.3.7).

Putting together the results of this section, we obtain the following.

**Proposition 3.3.5.** *For any $\varepsilon > 0$, we can find a function $F : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d \to [0, 1]$ which is Lipschitz, with*

$$t(F) \geqslant \gamma - \varepsilon$$

*where $t(F) = T(F)\Big/\big(\int F\big)^2$.*

*Proof.* Given $\varepsilon > 0$, choose $M$ large, and use Proposition 3.3.1 to find a set $S$ with $t(S) \geqslant \gamma - \varepsilon$, which is a subset of $[-M/2, M/2]$ of density at least $\eta$.

Next, choose an $\varepsilon'$ much smaller than $\varepsilon\eta^2$. Apply the arithmetic regularity lemma to the indicator function $f$ of $S$ to obtain a decomposition $f = f_{\mathrm{str}} + f_{\mathrm{sml}} + f_{\mathrm{unf}}$ as described in the statement of the arithmetic regularity lemma. The bounds (3.3.1) and (3.3.3) tell us that

$$\left|\frac{1}{M}\sum_n f(n) - \int F\right| \leqslant 2\varepsilon',$$

and the fact that $\frac{1}{M}\sum_n f(n) \geqslant \eta$ tells us that we must have

$$\int F \leqslant (1 + \varepsilon'/\eta)\frac{1}{M}\sum_n f(n).$$

Similarly, the bounds (3.3.2) and (3.3.7) tell us that

$$\left|\frac{1}{M^2}T(f) - T(F)\right| \leqslant 2\varepsilon',$$

and the fact that $T(f) \geqslant \frac{\eta^2}{13}$ tells us that

$$T(F) \geqslant (1 - 26\varepsilon'/\eta^2)T(f).$$

Putting these together, we deduce that

$$t(F) \geqslant \frac{1 - 26\varepsilon'/\eta^2}{1 + \varepsilon'/\eta}t(f).$$

Assuming that $\varepsilon'$ is small enough, we have that the fraction $\frac{1-26\varepsilon'/\eta^2}{1+\varepsilon'/\eta}$ is at least $1-\varepsilon$, and so $t(F) \geqslant \gamma - 2\varepsilon$. This completes the proof of Proposition 3.3.5. □

### 3.3.3 Simplifying the local problem

By Proposition 3.3.5, we know that $\gamma_{a_1,a_2,a_3}$ must be the supremum of $t(F)$ among functions $F$ defined on $\mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$. The goal of this section is to simplify this setting by showing that for any $x \in \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$, $F|_x$ must be the indicator function of an interval. Here, $F|_x$ represents the function defined on $[-1/2, 1/2]$ by $F|_x(y) = F(x_{\text{rat}}, y, x_{\text{irrat}})$, where $x_{\text{rat}} \in \mathbb{Z}/q\mathbb{Z}$ and $x_{\text{irrat}} \in \mathbb{T}^d$ are the components of $x$. Once we have done this, we will be able show that it suffices to consider functions with $d = 0$, and this will essentially be enough to establish Lemma 3.1.5. We will now begin with the first of these goals.

**Proposition 3.3.6.** *Let $F : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d \to [0, 1]$ be a Lipschitz function. Define $F^\# : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d \to [0, 1]$ to be the function with the following properties:*

- *For any $x \in \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$, $F^\#|_x : [-1/2, 1/2] \to [0, 1]$ is the indicator function of an interval centred at 0;*
- *For any $x \in \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$,*

$$\int_{-1/2}^{1/2} F|_x(t)dt = \int_{-1/2}^{1/2} F^\#|_x(t)dt.$$

*Then, $T(F) \leqslant T(F^\#)$, and thus $t(F) \leqslant t(F^\#)$.*

*Proof.* Let us begin by rewriting the $\mathbb{R}$ integral in (3.3.6). The integral operator may be written

$$\iint = \mathbb{E}_{u_{\mathrm{rat}},v_{\mathrm{rat}}\in\mathbb{Z}/q\mathbb{Z}} \int_{u_{\mathrm{irrat}},v_{\mathrm{irrat}}\in\mathbb{T}^d} \int_{u_{\mathrm{smth}},v_{\mathrm{smth}}\in\mathbb{R}}.$$

For fixed $u = (u_{\mathrm{rat}}, u_{\mathrm{irrat}})$ and $v = (v_{\mathrm{rat}}, v_{\mathrm{irrat}})$, the real integral becomes

$$\iint F_1(y_1u + z_1v)F_2(y_2u + z_2v)F_3(y_3u + z_3v)du_{\mathrm{smth}}dv_{\mathrm{smth}},$$

where $F_i = F|_{y_iu+z_iv}$. Now, we can certainly divide by $a_3$ in $\mathbb{R}$; we obtain

$$\iint \frac{1}{|a_3|}F_1(x_1)F_2(x_2)F_3(x_3)dx_1dx_2$$

where $x_3 = -\frac{a_1x_1+a_2x_2}{a_3}$, and the Jacobian factor of $\frac{1}{|a_3|}$ comes from (3.3.4).

In view of this, it will suffice to establish a version of Proposition 3.3.6 for separate functions $F_1, F_2, F_3$ defined on $[-1/2, 1/2]$. In particular, it will suffice to establish the following.

**Claim.** *For a function $F_i : [-1/2, 1/2] \to [0, 1]$, define $F_i^{\#}$ to be $(F^{\#})_i$. Then*

$$\iint F_1(x_1)F_2(x_2)F_3(x_3)dx_1dx_2 \leqslant \iint F_1^{\#}(x_1)F_2^{\#}(x_2)F_3^{\#}(x_3)dx_1dx_2 \tag{3.3.8}$$

*where the integrals run over $a_1x_1 + a_2x_2 + a_3x_3 = 0$.*

Now, [33, Theorem 379] gives us exactly that

$$\iint F_1(x_1)F_2(x_2)F_3(x_3)dx_1dx_2 \leqslant \iint F_1^{*}(x_1)F_2^{*}(x_2)F_3^{*}(x_3)dx_1dx_2, \tag{3.3.9}$$

where, for a function $f$, $f^*$ denotes the *symmetric rearrangement* of $f$; that is, the unique even function, decreasing on $\{x > 0\}$, with the property that $\mu(\{x : f(x) \geqslant t\}) = \mu(\{x : f^*(x) \geqslant t\})$ for each $t$ (where $\mu$ denotes the Lebesgue measure). An even function decreasing on $\{x > 0\}$ is called *symmetrically decreasing*.

Note the facts that

1. The convolution of symmetrically decreasing functions is symmetrically decreasing,
2. If $g$ and $h$ are symmetrically decreasing, then $\int gh \leqslant \int g^{\#}h$.

Given this, we can simply apply fact 2 to each of the functions $F_i^*$ on the right hand side of (3.3.9) in turn, to establish the claim. □

Now, we may rewrite $T(F)$. For $x \in \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$, let $f(x)$ be the length of the interval defined by $F|_x$. Then, we have that

$$\begin{aligned} T(F) &= \iint_{\mathbb{Z}/q\mathbb{Z}\times\mathbb{T}^d} \left( \frac{1}{|a_3|} \int_{t_1=-f(x_1)/2}^{f(x_1)/2} \int_{t_2=-f(x_2)/2}^{f(x_2)/2} \mathbb{1}_{|a_1t_1+a_2t_2|\leqslant a_3 f(3)/2|} dt_1 dt_2 \right) dudv \\ &= \iint_{\mathbb{Z}/q\mathbb{Z}\times\mathbb{T}^d} \left( \frac{1}{|a_1a_2a_3|} (I_{a_1f(x_1)} * I_{a_2f(x_2)} * I_{a_3f(x_3)})(0) \right) dudv \end{aligned}$$

where $x_i \in \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$ are defined by $x_i = y_iu + z_iv$, and recalling that $I_x$ is the indicator of an interval of length $x$ centred at 0.

**Definition 3.3.7.** For $x_1, x_2, x_3 \geqslant 0$, define $\Phi(x_1, x_2, x_3)$ to be the value of $I_{x_1} * I_{x_2} * I_{x_3}(0)$, where $*$ denotes convolution.

In view of Definition 3.3.7, we can now rewrite the definition of $T$. If $f$ is a function defined on $\mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d$ representing the length of the intervals (as above), then we have

$$T(f) = \frac{1}{|a_1a_2a_3|} \iint \Phi\left(a_1f(x_1), a_2f(x_2), a_3f(x_3)\right) dudv. \tag{3.3.10}$$

We will now state some useful properties of $\Phi$.

**Proposition 3.3.8.**

1. *Suppose that $b_1 \leqslant b_2 \leqslant b_3$. Then*

$$\Phi(b_1, b_2, b_3) = \begin{cases} b_1b_2 & b_3 \geqslant b_1 + b_2 \\ \dfrac{2(b_1b_2 + b_2b_3 + b_3b_1) - (b_1^2 + b_2^2 + b_3^2)}{4} & \text{otherwise.} \end{cases} \tag{3.3.11}$$

2. *For any $b_1, b_2$ and $b_3$,*

$$\Phi(b_1, b_2, b_3) \leqslant b_1b_2 \tag{3.3.12}$$

3. *For any $b_1, b_2, b_3 > 0$, there exist $p_1, p_2, p_3 > 0$ so that for any $x_1, x_2, x_3 > 0$,*

$$\Phi(x_1, x_2, x_3) \leqslant p_1x_2x_3 + p_2x_3x_1 + p_3x_1x_2 \tag{3.3.13}$$

   *with equality holding when $x_i = b_i$ for each $i$.*

4. *For any $b_1, b_2$ and $b_3$,*

$$\Phi(b_1, b_2, b_3) \leqslant \frac{1}{4}(b_1b_2 + b_2b_3 + b_1b_3) \tag{3.3.14}$$

5. *Let $G$ be any finite abelian group, and let $f_1, f_2, f_3 : G \to [0,1]$ with $b_i = \sum_{g\in G} f_i(g)$. Then*

$$\sum_{\substack{g_1,g_2,g_3\in G\\ g_1+g_2+g_3=0}} \Phi(f_1(g_1), f_2(g_2), f_3(g_3)) \leqslant \Phi(b_1, b_2, b_3) \\ = |G|^2\Phi(b_1/|G|, b_2/|G|, b_3/|G|) \tag{3.3.15}$$

*Proof.* Establishing (3.3.11) is fairly straightforward via direct integration, and equation (3.3.12) follows directly from that. Equation (3.3.13) generalises this; to prove it, it is easiest to prove the equivalent statement that the function

$$\Theta(x_1, x_2, x_3) = x_1x_2x_3\Phi(1/x_1, 1/x_2, 1/x_3)$$

is a concave function, which can be proven by showing that the Hessian matrix is always non-positive definite (for example, by using Sylvester's Rule). Equation (3.3.14) is then a special case of (3.3.13).

To establish (3.3.15), observe that the equality on the right hand side follows from the homogeneity of $\Phi$. The inequality immediately follows from applying (3.3.13) (with this choice of $b_1, b_2$ and $b_3$) to each term on the left hand side, observing that each pair $g_1, g_2 \in G$ occurs together in exactly one term on the left hand side (and likewise with pairs $g_1, g_3$ and $g_2, g_3$). □

*Remark.* Observe that (3.3.14) is a continuous analogue of (3.2.2).

With these, we can complete the proof of Lemma 3.1.5. We will first establish the following.

**Lemma 3.3.9.** *Let $G$ be any abelian group with a subgroup $H$. Suppose that $H$ is finite and has order coprime to $a_1a_2a_3$. Let $f : G \to [0,1]$ be any function, and define $f' : G \to [0,1]$ to be the function such that $f'(g) = \mathbb{E}_{h\in g+H} f(h)$. Then $T(f) \leqslant T(f')$.*

*Proof.* Write

$$\begin{aligned} T(f) &= \frac{1}{|a_1a_2a_3|}\iint \Phi(a_1f(x_1), a_2f(x_2), a_3f(x_3))dudv \\ &= \frac{1}{|a_1a_2a_3|}\iint \left(\mathbb{E}_{\substack{y_1,y_2,y_3\in H\\ \sum a_iy_i=0}} \Phi(a_1f(x_1+y_1), a_2f(x_2+y_2), a_3f(x_3+y_3))\right) dudv. \end{aligned}$$

For each $i$, set $f_i(h) = f(x_i + h/a_i)$ for $h \in H$, where we can divide by $a_i$ because $a_i$ is coprime to the order of $H$. Then, the integrand becomes

$$\mathbb{E}_{\substack{h_1,h_2,h_3\in H\\ h_1+h_2+h_3=0}} \Phi(a_1f_1(h_1), a_2f_2(h_2), a_3f_3(h_3)).$$

The inequality (3.3.15) immediately tells us that

$$\mathbb{E}_{y_1,y_2,y_3\in H}\Phi(a_1f_1(y_1),a_2f_2(y_2),a_3f_3(y_3)) \leqslant \Phi(a_1f'(x_1),a_2f'(x_2),a_3f'(x_3)).$$

In other words,

$$\begin{aligned} T(f) &\leqslant \frac{1}{|a_1a_2a_3|}\iint \Phi(a_1f'(x_1),a_2f'(x_2),a_3f'(x_3))dudv \\ &= T(f') \end{aligned}$$

as required. □

We are now ready to complete the proof of Lemma 3.1.5. We wish to show that the maximal value of $t(f)$ over all choices of $q \in \mathbb{N}$, $d \geqslant 0$ and $f : \mathbb{Z}/q\mathbb{Z} \times \mathbb{T}^d \to [0,1]$ is achieved when $d = 0$, and all prime factors of $q$ divide $a_i$ for some $i$.

First, observe that Lemma 3.3.9 immediately tells us that we can assume all prime factors of $q$ divide some $a_i$, by taking $H \leqslant G$ to be the subgroup of $\mathbb{Z}/q\mathbb{Z}$ consisting of elements coprime to each $a_i$. We just need to establish that $d$ can be taken to be 0. This will be made possible via the following claim:

**Claim.** *If $f : G \times \mathbb{T} \to [0,1]$ is any Lipschitz function for an abelian group $G$, and $g : G \to [0,1]$ is defined by $g(x) = \mathbb{E}_y f((x,y))$, then $t(f) \leqslant t(g)$.*

*Proof of Claim.* Let $p$ be some prime that is coprime to each $a_i$. Given $f$, define $f_n$ to be the function obtained as follows. Take $H \leqslant G \times \mathbb{T}$ to be the subgroup of $\mathbb{T}$ of order $p^n$. Let $f_n$ be the function obtained by applying Lemma 3.3.9 to $f$ and $H$. Thus, $f = f_0$ and $T(f_i) \leqslant T(f_j)$ for each $i < j$.

The claim will follow from the assertion that $f_n$ converges to a function which is constant on $\mathbb{T}$, and this follows from the Lipschitz property. ◊

Finally, to complete the proof of Lemma 3.1.5, we need only observe that, if we are given a number $q$ and a function $f : \mathbb{Z}/q\mathbb{Z} \to [0,1]$ (which we may normalise so that $\sum f = 1$), then $t(S_M) \to t(f)$ as $M \to \infty$ with $S_M$ defined as in (3.1.5).

*Remark.* This proof of Lemma 3.1.5 can also be used to establish the remark after Definition 3.1.3 regarding replacing lim sup with lim. Indeed, we can show that $\max_{|S|=n} \frac{T(S)}{|S|^2}$ converges as $n \to \infty$ as follows. For any $\varepsilon > 0$, there must exist a choice of $q \in \mathbb{N}$ and $f : \mathbb{Z}/q\mathbb{Z} \to [0,1]$ such that $t(f) > \gamma_{a_1,a_2,a_3} - \varepsilon$. Now, for any $M$, the set $S_M$ satisfies $|S_M| = M + O(q)$. In particular, we have that $t(S_M) > \gamma - 2\varepsilon$ for all $M$. Since $|S_M|$ ranges over a set of values with bounded gaps, it must be the case that for all sufficiently large $M$, there is a set $S'_M$ of size $M$ with the property that $t(S'_M) > \gamma - 3\varepsilon$. This yields the claimed result.

## 3.4 Completing the proof of Theorem 3.1.4

In this section, we will show how Lemma 3.1.5 may be used to prove Theorem 3.1.4.

By Lemma 3.1.5, we need only consider functions $f$ defined on $\mathbb{Z}/q\mathbb{Z}$, with $T(f)$ defined as in (3.3.10). Indeed, we can write

$$T(f) = \sum_{\substack{x_1,x_2,x_3 \in \mathbb{Z}/q\mathbb{Z} \\ a_1x_1+a_2x_2+a_3x_3=0}} \frac{1}{|a_1a_2a_3|}\Phi(a_1f(x_1), a_2f(x_2), a_3f(x_3)) \tag{3.4.1}$$

We will start with the following lemma.

**Lemma 3.4.1.** *Fix $\delta > 0$ small. Say that a term $\Phi(a_1f(x_1), a_2f(x_2), a_3f(x_3))$ in (3.4.1) is* bad *if there is a pair $i, j$ for which $\frac{a_if(x_i)}{a_jf(x_j)} \leqslant \frac{\delta}{6}$.*

*Then the contribution to $T(f)$ from the bad pairs is no more than $\delta\left(\sum_x f(x)\right)^2$; in other words, the contribution to $t(f)$ is no more than $\delta$.*

*Proof.* It will suffice to show that the contribution to $T(f)$ from terms with $\frac{a_1f(x_1)}{a_2f(x_2)} \leqslant \delta$ is no more than $\delta\left(\sum_x f(x)\right)^2$. And,

$$\begin{aligned}
\sum_{\substack{a_1x_1+a_2x_2+a_3x_3=0 \\ a_1x_1 \leqslant \delta a_2x_2}} \frac{1}{a_1a_2a_3}\Phi(a_1f(x_1), a_2f(x_2), a_3f(x_3)) &\leqslant \sum_{\substack{a_1x_1+a_2x_2+a_3x_3=0 \\ a_1x_1 \leqslant \delta a_2x_2}} \frac{1}{a_2}f(x_1)f(x_3) \\
&\leqslant \sum_{\substack{a_1x_1+a_2x_2+a_3x_3=0 \\ a_1x_1 \leqslant \delta a_2x_2}} \frac{\delta}{a_1}f(x_2)f(x_3) \\
&\leqslant \sum_{x_2,x_3} \delta f(x_2)f(x_3) \\
&= \delta\left(\sum_x f(x)\right)^2,
\end{aligned}$$

where the first inequality follows from (3.3.12), and the third follows from the fact that because $\mathbb{Z}/q\mathbb{Z}$ is cyclic, there can be at most $a_1$ solutions to $a_1x_1+a_2x_2+a_3x_3 = 0$ for $x_2$ and $x_3$ fixed. □

*Proof of Theorem 3.1.4.* Given $\varepsilon > 0$ small, let $\beta = \frac{6}{\varepsilon}$. By Lemma 3.4.1, the contribution to $T(f)$ from terms for which at least one of the ratios between two of the $a_if(x_i)$ is smaller than $1/\beta$ is at most $\varepsilon$. We just need to show that, for a judicious choice of $a_1, a_2$ and $a_3$, the contribution from the remaining terms must be at most $\frac{1}{12}$. Our choice will be $a_1 = 1, a_2 \approx \beta^r$ and $a_3 \approx \beta^{r+s}$, with $r = 10^3$ and $s = 10^6$, where we insist that the $a_i$ are integers.

We will bound the sum of the good terms $\Phi(a_1 f(x_1), a_2 f(x_2), a_3 f(x_3))$ by using (3.3.14). We obtain

$$\sum_{\substack{a_1x_1+a_2x_2+a_3x_3=0 \\ \text{term is good}}} \frac{1}{a_1a_2a_3}\Phi(a_1f(x_1), a_2f(x_2), a_3f(x_3))$$
$$\leqslant \frac{1}{4} \sum_{\substack{a_1x_1+a_2x_2+a_3x_3=0 \\ (x_1,x_2,x_3) \text{ is good}}} \frac{f(x_1)f(x_2)}{a_3} + \frac{f(x_1)f(x_3)}{a_2} + \frac{f(x_2)f(x_3)}{a_1}. \quad (3.4.2)$$

Select a pair of elements $u, v$ in $\mathbb{Z}/q\mathbb{Z}$ with $f(u)$ and $f(v)$ nonzero. Suppose that $\frac{a_1 f(u)}{a_2 f(v)}$ is between $1/\beta$ and $\beta$; in other words, there may exist choices of $w$ for which $(u, v, w)$ is a good solution to $a_1u+a_2v+a_3w = 0$. Then, by our choice of $a_1, a_2$ and $a_3$, none of the other five ratios $\frac{a_i f(u)}{a_j f(v)}$ can be between $1/\beta$ and $\beta$.

Now, let us bound the coefficient of $f(u)f(v)$ in the right hand side of (3.4.2). There can be at most $a_3$ choices of $w$ for which $a_1u + a_2v + a_3w = 0$, and so $\frac{1}{4}\frac{f(u)f(v)}{a_3}$ can appear at most $a_3$ times in the sum. The same argument applies for each of the other five ratios that could be between $1/\beta$ and $\beta$. Therefore, for any pair $u$ and $v$, the coefficient of $f(u)f(v)$ in (3.4.2) will be at most $\frac{1}{4}$.

Thus,

$$T(f) \leqslant \sum_{\substack{x_1,x_2 \in \mathbb{Z}/N\mathbb{Z} \\ x_1,x_2 \text{ valid}}} \frac{1}{4} f(x_1)f(x_2), \quad (3.4.3)$$

where $x_1, x_2$ is *valid* if they are in at least one good triple (and thus one of the six ratios is between $\frac{1}{\beta}$ and $\beta$).

Now, partition the elements $x \in \mathbb{Z}/q\mathbb{Z}$ for which $f(x) \neq 0$ into sets $B_i$ for $i \in \mathbb{Z}$, where $x \in B_i$ if $\beta^i \leqslant f(x) < \beta^{i+1}$. Let $S_i = \sum_{x \in B_i} f(x)$. The right hand side of (3.4.2) is bounded by

$$\frac{1}{4} \sum_{\substack{i<j \in \mathbb{Z} \\ j-i \text{ valid}}} S_iS_j \quad (3.4.4)$$

where $j - i$ is valid if and only if it differs from one of $\{r, s, r + s\}$ by at most 1.

We will visualise this as a graph $G$ whose vertices are the set of $i$ for which $S_i$ is nonzero, and with an edge connecting $i$ and $j$ if and only if $j - i$ is valid. Then, the following three claims complete the proof of the theorem.

**Claim.** *Suppose that $G$ is 3-colourable. Then the quantity (3.4.4) is not greater than $\frac{1}{12}\left(\sum_i S_i\right)^2$.*

*Proof of Claim.* 3-colour $G$ and let $T_1, T_2$ and $T_3$ be the sums of $S_i$ over the three colour classes. Then

$$\begin{aligned}\frac{1}{4}\sum_{\substack{i<j\in\mathbb{Z}\\ j-i \text{ valid}}} S_iS_j &\leqslant \frac{1}{4}\left(T_1T_2+T_2T_3+T_3T_1\right)\\ &\leqslant \frac{1}{12}\left(T_1+T_2+T_3\right)^2\\ &= \frac{1}{12}\left(\sum_i S_i\right)^2,\end{aligned}$$

where the second inequality follows from an application of the Cauchy-Schwarz inequality. ◊

**Claim.** *If $G$ has fewer than 500 vertices, then it must be 3-colourable.*

*Proof of Claim.* Without loss of generality, $G$ is connected and contains 0 as one of its vertices (else we may consider a connected component of $G$ and translate the vertices). Since $G$ is connected, any vertex may be written as $a+rb+sc$, for some $|a|,|b|,|c|<500$; note that this bound follows because along each edge, $a, b$ and $c$ can each change by at most 1. By our choice of $r=10^3$ and $s=10^6$, this representation must be unique. Thus, we may colour the vertices of $G$ according to $b+c \bmod 3$. ◊

**Claim.** *The quantity*

$$\left(\sum_{\substack{i<j\in\mathbb{Z}\\ j-i \text{ valid}}} S_iS_j\right)\Big/\left(\sum_i S_i\right)^2 \tag{3.4.5}$$

*achieves its maximum at a point where the number of nonzero $S_i$ is less than 500.*

*Proof of Claim.* First, observe that, by homogeneity, we may restrict our attention to functions with $S=\sum_i S_i = 1$. The set of functions with support of size at most $N$ and sum 1 is compact, so (3.4.5) achieves its maximum.

At the point where (3.4.5) is maximal, we can see that for any $i$ and $j$ for which $S_i, S_j \neq 0$, the derivatives of $\sum_{\substack{i<j\in\mathbb{Z}\\ j-i \text{ valid}}} S_iS_j$ with respect to $S_i$ and $S_j$ must be equal. Let that quantity be $M$, so

$$M = \sum_{\substack{j\in\mathbb{Z}\\ |j-i| \text{ valid}}} S_j.$$

We may rewrite (3.4.5) as $\frac{1}{2}\frac{SM}{S^2} = \frac{M}{2}$, and so at the maximum $M \geqslant \frac{2}{3}$. Since each vertex of $G$ has degree at most 18, we deduce that whenever $S_i$ is nonzero, there is a $j$ connected to $i$, with $S_j \geqslant \frac{1}{27}$.

There are at most 27 choices of $j$ with $S_j \geqslant \frac{1}{27}$, and each of those can be connected to at most 18 choices of $i$, and so there can be at most 486 nonzero $S_i$. $\Diamond$

The three claims prove that the quantity (3.4.5) must be bounded above by $\frac{1}{12}$. Theorem 3.1.4 follows immediately. $\square$

## 3.5 Equations in more than three variables

A fairly natural extension of Theorem 3.1.4 is to ask if a similar result holds for $k$-variable equations

$$a_1 x_1 + \cdots + a_k x_k = 0. \tag{3.5.1}$$

As before, let $T(S)$ be the number of solutions to (3.5.1) in $S$. Similarly, denote the number of solutions with $x_i$ taken from $S_i$ by $T(S_1, \ldots, S_k)$. We have a trivial upper bound for $T(S)$, namely that $T(S) \leqslant |S|^{k-1}$.

For an integer $k \geqslant 3$, define

$$\sigma_k = (\underbrace{I_1 * \cdots * I_1}_{k})(0), \tag{3.5.2}$$

recalling that $I_1$ denotes the indicator function of an interval of length 1. In particular, $\sigma_3 = 3/4$.

*Remark.* We also have another simple formula for $\sigma_k$:

$$\sigma_k = \int_{-\infty}^{\infty} \left( \frac{\sin \pi x}{\pi x} \right)^k dx. \tag{3.5.3}$$

The equivalence of these forms follows from taking a Fourier transform and applying the convolution identity; the details can be seen in [8]. See also [46], where it can be shown that $\sigma_{2h}$ is the leading coefficient of the polynomial $\Psi_h(n)$ (using the notation of that paper).

*Remark.* $\sigma_k$ obeys a simple asymptotic (see for example [47], or [23] for more terms):

$$\sigma_k = \sqrt{\frac{6}{k\pi}} \, (1 + O(1/k))$$

as $k \to \infty$.

We may interpret $\sigma_k$ combinatorially. If $f_k$ is the probability density function of a sum of $k$ independent random variables distributed uniformly on $[-1/2, 1/2]$, then $\sigma_k = f_k(0)$. Thus, the form of the asymptotic for $\sigma_k$ is not surprising, in view of the Central Limit theorem.

We may also generalise $\Phi$ as defined by Definition 3.3.7:

**Definition 3.5.1.** For $t_1, \ldots, t_k$ positive real numbers, define the function $\Phi = \Phi_k$ by

$$\Phi_k(t_1, \ldots, t_k) = (I_{t_1} * \cdots * I_{t_k})(0). \tag{3.5.4}$$

In particular, $\sigma_k = \Phi_k(1, \ldots, 1)$.

*Remark.* There is an explicit formula for $\Phi_k$. In general, we have

$$\Phi_k(t_1, \ldots, t_k) = \frac{1}{(k-1)!2^k} \sum_{\varepsilon \in \{\pm 1\}^k} \omega(\varepsilon)(\varepsilon \cdot \mathbf{t})^{k-1} \operatorname{sgn}(\varepsilon \cdot \mathbf{t}), \tag{3.5.5}$$

where $\omega(\varepsilon) = \prod_i \varepsilon_i$, $\varepsilon \cdot \mathbf{t} = \sum_i \varepsilon_i t_i$ and sgn denotes the sign function. This is established in [8].

Also, note that this reduces to (3.3.11) when $k = 3$.

By analogy with Proposition 3.1.2, we have the following.

**Proposition 3.5.2.** *Let $k \geqslant 3$ be an integer. For any equation $a_1x_1 + \cdots + a_kx_k = 0$, there are large sets $S$ for which*

$$T(S) \geqslant \frac{\sigma_k}{k^{k-1}}|S|^{k-1} + O(|S|^{k-2}). \tag{3.5.6}$$

The proof of Proposition 3.5.2 will rely on the following fact, which states that, when the coefficients $a_i$ are all 1, long progressions behave somewhat like real intervals.

**Proposition 3.5.3.** *Suppose $S_1, \ldots, S_k$ are arithmetic progressions centred at the origin, with the same common difference. Let $s_i$ be the number of terms in $S_i$.*

*Then, the number of solutions to $x_1 + \cdots + x_k = 0$ where each $x_i \in S_i$ is*

$$\Phi(s_1, \ldots, s_k) + O_k((s_1 + \cdots + s_k)^{k-2}).$$

*Proof.* We may assume without loss of generality that the progressions $S_i$ have common difference 1. To prove Proposition 3.5.3, it suffices to use the following observation.

Suppose that $y_1, \ldots, y_{k-1}$ are elements of the real intervals $I_{s_1}, \ldots, I_{s_{k-1}}$. Then, we have the following two implications for $k-1$-tuples of real numbers $y_1, \ldots, y_{k-1}$.

- If $|y_1 + \cdots + y_{k-1}| \leqslant \frac{s_k}{2}$, then $|\lfloor y_1 \rfloor + \cdots + \lfloor y_{k-1} \rfloor| \leqslant \frac{s_k}{2} + k$;
- If $|\lfloor y_1 \rfloor + \cdots + \lfloor y_{k-1} \rfloor| \leqslant \frac{s_k}{2} - k$, then $|y_1 + \cdots + y_{k-1}| \leqslant \frac{s_k}{2}$.

Now, $T(S_1, \dots, S_k)$ counts the number of $k-1$-tuples of integers $(x_i)_{i=1}^{k-1}$ with $x_i \in S_i$, such that $-\sum_i x_i \in S_k$.

Up to an error which is at most $O_k((s_1 + \dots + s_k)^{k-2})$, this can be written as an integral

$$\int_{-s_1/2}^{s_1/2} \dots \int_{-s_{k-1}/2}^{s_{k-1}/2} \mathbb{1}_{|\lfloor y_1 \rfloor + \dots + \lfloor y_{k-1} \rfloor| \leqslant s_k/2} dy_1 \dots dy_{k-1}.$$

The two implications above allow us to show that, up to acceptable error, this is equal to

$$\int_{-s_1/2}^{s_1/2} \dots \int_{-s_{k-1}/2}^{s_{k-1}/2} \mathbb{1}_{|y_1 + \dots + y_{k-1}| \leqslant s_k/2} dy_1 \dots dy_{k-1},$$

which is equal to $\Phi(s_1, \dots, s_k)$. □

We are now ready to prove Proposition 3.5.2.

*Proof of Proposition 3.5.2.* As in Proposition 3.1.2, we will consider $S$ as the union of $k$ sets $S_1, \dots, S_k$, with the property that $T(S_1, \dots, S_k)$ is large.

The way we will do this is as follows. Let $M$ be a large integer, which we assume to be divisible by $2k$. Define

$$S_i = \frac{1}{a_i}[-M/2k, M/2k]$$

for each $i$ with $1 \leqslant i \leqslant k$, where we may normalise the sets to consist of integers by multiplying by $\prod_i a_i$. Then, let $S = \cup_i S_i$, so that $|S| \leqslant M$.

It remains to show that $T(S_1, \dots, S_k) \geqslant \frac{\sigma_k}{k^{k-1}} M^{k-1} + O(M^{k-2})$. But this follows as an easy consequence of Proposition 3.5.3. Indeed,

$$\begin{aligned} T(S_1, \dots, S_k) &= \Phi_k(M/k, \dots, M/k) + O(M^{k-2}) \\ &= \left(\frac{M}{k}\right)^{k-1} \Phi_k(1, \dots, 1) + O(M^{k-2}) \\ &= \frac{\sigma_k}{k^{k-1}} M^{k-1} + O(M^{k-2}). \end{aligned}$$

□

Perhaps unsurprisingly, Theorem 3.1.4 also generalises to this setting.

**Theorem 3.5.4.** *Let $\varepsilon > 0$. Then, there exist coefficients $a_1, \dots, a_k$ with the property that, for any suitably large set $S$,*

$$T(S) \leqslant \left(\frac{\sigma_k}{k^{k-1}} + \varepsilon\right) |S|^{k-1} + o(|S|^{k-1}).$$

In this section, we will discuss how to generalise the elementary proof from Section 3.2 to prove Theorem 3.5.4. There are two main places in which the argument slightly differs from the argument given in Section 3.2. Firstly, we must generalise Lemma 3.2.3 to give a bound for $T(S_1, \dots, S_k)$ in terms of $E(S_1, S_2)$:

**Lemma 3.5.5.** *Suppose that $S_1, \dots, S_k \subseteq \mathbb{Z}$ are finite sets. Then*

$$T(S_1, \dots, S_k)^2 \leqslant E(a_1 \cdot S_1, a_2 \cdot S_2) \left(|S_3||S_4| \dots |S_k|\right)^{2-1/(k-2)}.$$

*Proof.* For any $t \in \mathbb{Z}$, let $\mu(t)$ denote the number of ways of writing $t = a_1x_1 + a_2x_2$, for $x_i \in S_i$. Thus, by definition,

$$E(a_1 \cdot S_1, a_2 \cdot S_2) = \sum_t \mu(t)^2.$$

Define $\nu(t)$ to be the number of ways of writing $t = -a_3x_3 - \dots - a_kx_k$, for $x_i \in S_i$. Thus, we see that

$$\begin{aligned} T(S_1, \dots, S_k)^2 &= \left(\sum_t \mu(t)\nu(t)\right)^2 \\ &\leqslant \left(\sum_t \mu(t)^2\right)\left(\sum_t \nu(t)^2\right) \\ &= E(a_1 \cdot S_1, a_2 \cdot S_2)\left(\sum_t \nu(t)^2\right). \end{aligned}$$

Finally, we observe that $\sum_t \nu(t)^2$ represents the number of solutions to the equation

$$a_3x_3 + \dots + a_kx_k = a_3x_3' + \dots + a_kx_k',$$

and so we can bound it by $(|S_3| \dots |S_k|)^{2-\frac{1}{k-2}}$, by the same argument used in (3.2.1) to bound the energy. □

Secondly, we will have to apply a $k$ variable analogue of Lemma 3.2.6. The analogue of this is the following:

**Lemma 3.5.6.** *Suppose that $S_1, \dots, S_k \subseteq \mathbb{Z}$ are sets with $|S_i| = s_i$. Then*

$$T(S_1, \dots, S_k) \leqslant \frac{\sigma_k}{k}\left(\sum_i \hat{s_i}\right) + O\left(\sum_i s_i\right)^{k-2},$$

*where $\hat{s_i} = \prod_{j \neq i} s_j$.*

*Remark.* This lemma is actually weaker than Lemma 3.2.6, where the error term was $O(1)$. The weaker error term here comes from our reduction to the real case using Proposition 3.5.3; an inductive proof would likely give an $O(\sum s_i)^{k-3}$ error term. However, the $O(\sum s_i)^{k-2}$ error term is sufficient for our purpose.

*Proof of Lemma 3.5.6.* First, observe that the statement of the lemma is unchanged if we assume without loss of generality that each $a_i$ is 1, since we may replace $S_i$ with $a_i \cdot S_i$.

The first step in the proof is to apply [43, Theorem 1], which says that we may take each $S_i$ to be an interval of length $s_i$, roughly centred at the origin (depending on the parity of $s_i$), in order to maximise $T_{1,\dots,1}(S_1, \dots, S_k)$. We may immediately apply Proposition 3.5.3, which says that

$$T_{1,\dots,1}(S_1, \dots, S_k) = \Phi_k(s_1, \dots, s_k) + O((s_1 + \cdots + s_k)^{k-2}).$$

Thus, it suffices to prove that

$$\Phi(s_1, \dots, s_k) \leqslant \frac{\sigma_k}{k} \left( \sum_i \hat{s_i} \right). \tag{3.5.7}$$

This will follow if we can prove that, for positive real numbers $t_1, \dots, t_k$,

$$t_1 \dots t_k \Phi(t_1^{-1}, \dots, t_k^{-1}) \leqslant \frac{\sigma_k}{k}(t_1 + \cdots + t_k). \tag{3.5.8}$$

To prove (3.5.8), first observe that equality holds in the case that all of the $t_i$ are equal. Indeed, when $t_i = 1$ the relation follows from the definition of $\sigma$, and for other constant values of $t_i$ the equality follows by homogeneity.

Set $\Theta(t_1, \dots, t_k) = t_1 \dots t_k \Phi(t_1^{-1}, \dots, t_k^{-1})$. To prove that $\Theta(t_1, \dots, t_k)$ achieves its maximum value (with $t_1 + \cdots + t_k$ fixed) when all of the $t_i$ are equal, observe that it will suffice to prove the following claim.

**Claim 1.** *If $t_1 + t_2$ is fixed (as well as each of $t_3, \dots, t_k$), then $\Theta(t_1, \dots, t_k)$ achieves its maximum when $t_1 = t_2$.*

To see that this claim is sufficient, observe that we may repeatedly replace the largest and smallest of the $t_i$ with their average. In doing so, $\max t_i - \min t_i$ will tend to 0, and we can use the continuity of $\Theta$ to obtain the result.

To prove Claim 1, recall the expression for $\Theta(t_1, \dots, t_k)$:

$$\begin{aligned}\Theta(t_1, \dots, t_k) &= t_1 t_2 (t_3 \dots t_k)(I_{t_1^{-1}} * I_{t_2^{-1}}) * (I_{t_3^{-1}} * \cdots * I_{t_k^{-1}})(0) \\ &= (t_1 t_2 (I_{t_1^{-1}} * I_{t_2^{-1}}) * g)(0),\end{aligned}$$

where $g(x) = t_3 \dots t_k (I_{t_3^{-1}} * \cdots * I_{t_k^{-1}})(x)$.

Now, observe that $g$ may be written as a combination of intervals, in the following sense:

$$g(x) = \int_0^\infty h(r) I_r(x) dr,$$

for some function $h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ with bounded support. (The exception is when $k = 3$, in which case $g$ is just a single interval. But that will not affect the remainder of the proof of Claim 1.)

To see why this is the case, we may use induction. If $k = 4$, then suppose without loss of generality that $t_3 \leqslant t_4$. Then, we take $h(r) = t_3 t_4$ if $\frac{t_3^{-1} - t_4^{-1}}{2} \leqslant r \leqslant \frac{t_3^{-1} + t_4^{-1}}{2}$, and 0 otherwise. For $k > 4$, it is easiest to apply the induction hypothesis to $I_{t_3^{-1}} * \cdots * I_{t_{k-1}^{-1}}$, and then use a similar decomposition to the one we used for the $k = 4$ case.

In view of this decomposition, proving Claim 1 may be reduced to the following claim:

**Claim 2.** *Fix $t_1 + t_2$. Then, for any choice of $t$, we have that $t_1 t_2 (I_{t_1^{-1}} * I_{t_2^{-1}} * I_t)(0)$ is maximised when $t_1 = t_2$.*

To prove Claim 2, we may use (3.3.13). Indeed, choose $p_1, p_2$ and $p_3$ such that equality occurs in the relation $\Phi(x_1, x_2, x_3) \leqslant p_1 x_2 x_3 + p_2 x_3 x_1 + p_3 x_1 x_2$ when $x_1 = x_2 = (t_1 + t_2)/2$ and $x_3 = t$; by symmetry, we will have $p_2 = p_3$. Then, we have

$$\begin{aligned} t_1 t_2 (I_{t_1^{-1}} * I_{t_2^{-1}} * I_t)(0) &= t_1 t_2 \Phi(t_1^{-1}, t_2^{-1}, t) \\ &\leqslant p_1 + p_2 t (t_1 + t_2) \end{aligned}$$

which is constant (as $t_1$ and $t_2$ vary). Therefore, the left hand side is maximised when $t_1 = t_2$. □

*Remark.* If $k$ is even, we can actually deduce a stronger version of (3.5.7) by using Hölder's inequality. We have

$$\begin{aligned} \Phi_k(s_1, \dots, s_k) &= \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{I_{s_1}}(r) \dots \widehat{I_{s_k}}(r) dr \\ &\leqslant \frac{1}{2\pi} \left( \prod_i \int_{-\infty}^\infty \widehat{I_{s_i}}(r) dr \right)^{1/k} \\ &= \prod_i \Phi_k(s_i, \dots, s_i)^{1/k} \\ &= \sigma_k(s_1 \dots s_k)^{1 - 1/k}, \end{aligned}$$

where the second line used Hölder's inequality along with the fact that $k$ is even. This is stronger than (3.5.7) via an application of the AM-GM inequality.

It is unclear whether the stronger version holds in the case that $k$ is odd; indeed, it is not too hard to establish for $k = 3$ by using (3.3.11). However, this stronger form is not necessary, so we only prove the version we need.

Armed with our more general Lemmas 3.5.5 and 3.5.6, we may use an argument similar to the proof of Theorem 3.1.4 in section 3.2 in order to prove Theorem 3.5.4.

*Sketch proof of Theorem 3.5.4.* Select $a_1, \dots, a_k$ to be coprime integers so that $a_1 = 1$, and $|a_i|$ is sufficiently large for $i \neq 1$. Let $S$ be a large set of integers.

With Lemma 3.5.5 replacing Lemma 3.2.3, much of the argument is the same as the proof of Theorem 3.1.4:

- We start by using Theorem 3.2.2 to split $S = S_1 \amalg \dots \amalg S_n \amalg S_0$, and show that $S_0$ can be ignored.
- We can define the labelled digraph $G$ which captures almost all of the solutions to $a_1x_1 + \dots + a_kx_k = 0$.
- We can prove that the product of the labels along a cycle must be 1, allowing us to define the function $d : [n] \to \mathbb{Q}$ as in (3.2.8).
- This allows us to show that an upper bound for the number of solutions coming from good $k$-tuples is
$$\sum_d T(R_{d/a_1}, \dots, R_{d/a_k}),$$
where $R_d$ is the union of the $S_i$ with $d(i) = d$ (as in the case $k = 3$, this sum can have no more than $n$ terms).

Lemma 3.5.6 allows us to bound this:

$$\begin{aligned}\sum_d T(R_{d/a_1}, \dots, R_{d/a_k}) &\leqslant \sum_d \frac{\sigma_k}{k} \sum_i \widehat{R_{d/a_i}} + O(K_1|S|^{k-2}) \\ &\leqslant \frac{\sigma_k}{k} \sum_{(d_1,\dots,d_{k-1})} |R_{d_1}| \dots |R_{d_{k-1}}| + O(K_1|S|^{k-2}). \qquad (3.5.9)\end{aligned}$$

On the first line, $\widehat{R_{d/a_i}}$ denotes the product of the other $|R_{d/a_j}|$, and the error term comes from the fact that there are at most $n < K_1$ terms in the sum on the left hand side. On the second line, the sum is over unordered $k-1$-tuples $(d_1, \dots, d_{k-1})$ for which, for some ordering of the $a_i$, we have that $a_{i_1}d_1 = \dots = a_{i_{k-1}}d_{k-1}$; there can only be one such ordering by coprimality.

Now, for $i = 0, 1, \ldots, k-1$, define the quantity $X_i$ by

$$X_i = \sum_{\substack{d = a_2^{e_2} \ldots a_k^{e_k} \\ e_2 + 2e_3 + \cdots + (k-1)e_k \equiv d \mod k}} |R_d|.$$

Note that if $d$ is such that $R_d$ is nonempty, then the representation of $d$ as a product $d = a_2^{e_2} \ldots a_k^{e_k}$ exists due to how we constructed the labels, and is unique due to the coprimality of the $a_i$.

Now, suppose that $R_{d_1}$ and $R_{d_2}$ appear together in at least one term on the right hand side of (3.5.9). Then, $d_1/d_2 = a_i/a_j$ for some $i \neq j$, and so $R_{d_1}$ and $R_{d_2}$ contribute to different $X_i$.

Thus, we may upper bound the sum in the right hand side of (3.5.9):

$$\sum_{(d_1, \ldots, d_{k-1})} |R_{d_1}| \ldots |R_{d_{k-1}}| \leqslant \sum_i \widehat{X_i}, \tag{3.5.10}$$

where $\widehat{X_i} = \prod_{j \neq i} X_j$. This bound follows from the fact that each unordered $k-1$-tuple $|R_{d_1}| \ldots |R_{d_{k-1}}|$ on the left hand side contributes to exactly one of the terms on the right hand side.

We finish by applying the following claim.

**Claim.** *If $X_1, \ldots, X_k$ are nonnegative real numbers, then we have the bound*

$$\sum_i \widehat{X_i} \leqslant \frac{1}{k^{k-2}} \left( \sum_i X_i \right)^{k-1}. \tag{3.5.11}$$

*Proof.* Fix $(\sum_i X_i)$, so the right hand side of (3.5.11) is constant. By compactness, the left hand side achieves its maximum. Suppose that, at the maximum, some $X_i$ is equal to 0; without loss of generality, assume $X_k = 0$. Then the left hand side becomes $\prod_{i<k} X_i$, and the inequality follows from the AM-GM inequality and the fact that $\frac{1}{(k-1)^{k-1}} \leqslant \frac{1}{k^{k-2}}$.

Otherwise, suppose that maximum occurs when $X_u \neq X_v$, for some $1 \leqslant u < v \leqslant k$ and none of the $X_i$ are zero. Replacing $X_u$ and $X_v$ with $\frac{X_u + X_v}{2}$ will leave $\widehat{X_u} + \widehat{X_v}$ unchanged, and will increase the value of $X_u X_v$ and so increase the value of the left hand side. Thus, it is impossible for the maximum to be achieved when the $X_i$ are not all equal.

Equality occurs in (3.5.11) when all of the $X_i$ are equal, and the claim follows. $\Diamond$

Putting the claim together with our upper bound (3.5.10) above, we deduce that

$$\sum_d T(R_{d/a_1}, \ldots, R_{d/a_k}) \leqslant \frac{\sigma_k}{k^{k-1}} |S|^{k-1} + O(|S|^{k-2}),$$

which gives the bound in the statement of Theorem 3.5.4 when $|S|$ is large enough. $\square$

## 3.6 Generalising Lemma 3.1.5 to more than three variables

Another strategy we might wish to take to prove a result like Theorem 3.5.4 would be to prove a $k$ variable analogue of Lemma 3.1.5.

**Lemma 3.6.1.** *Suppose $\gamma_{a_1,\dots,a_k}$ is defined by*

$$\gamma_{a_1,\dots,a_k} = \limsup_{|S|\to\infty} \frac{T(S)}{|S|^{k-1}}.$$

*Then, in order to compute $\gamma$ it suffices to consider sets as in the statement of Lemma 3.1.5.*

*In particular,*

$$\gamma = \sup_{\substack{f:\mathbb{Z}/q\mathbb{Z}\to\mathbb{R}^{\geqslant 0} \\ \sum f=1}} \sum_{\sum a_i x_i = 0} \frac{1}{|a_1 \dots a_k|} \Phi(a_1 f(x_1), \dots, a_k f(x_k)),$$

*where the supremum runs over all choices of $q$, all of whose prime factors divide at least one of the $a_i$.*

Indeed, armed with Lemma 3.6.1, an argument almost identical to that given in Section 3.4, replacing (3.3.14) with its generalisation (3.5.7), will prove a version of Theorem 3.5.4 (with a weaker error term).

In this section, we will discuss where the proof of Lemma 3.6.1 differs from the proof of Lemma 3.1.5.

Recall the structure of the original proof of Lemma 3.1.5. We began by showing in Proposition 3.3.1 that, if $S$ is a set with $T(S)$ large, we find a set $S'$ with about the same number of solutions, but which is a dense subset of an interval $[-M/2, M/2]$.

Having done this, we could apply the arithmetic regularity lemma to $S'$. After some manipulation, this gave us Proposition 3.3.5, which essentially said that to calculate $\gamma_{a_1,a_2,a_3}$, it suffices to consider functions $F : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$.

Finally, we completed the proof of Lemma 3.1.5. We did this in two steps. First, we manipulated the expression for $t(F)$ to show that we could assume that, when we fix the components of $x$ in $\mathbb{Z}/q\mathbb{Z}$ and $\mathbb{T}^d$, the resulting function on $[-1/2, 1/2]$ is the indicator function of an interval centred at 0. This formed Proposition 3.3.6. Having done this, we used some concavity properties of the function $\Phi$ to complete the proof of Lemma 3.1.5.

When generalising this argument, the first two steps remain mostly unchanged. Indeed, the only major difference is that, instead of using the energy bound of Lemma

3.2.3, we use its $k$-variable analogue Lemma 3.5.5. This again reduces to the consideration of functions $F : \mathbb{Z}/q\mathbb{Z} \times [-1/2, 1/2] \times \mathbb{T}^d$. Furthermore, the argument in Proposition 3.3.6 carries through to the $k$-variable case exactly.

This leaves us with the completion of the proof of the Lemma. Unfortunately, this required the use of concavity properties of the function $\Phi$, which do not obviously generalise to the $k$ variable case.

In particular, we used (3.3.15), whose proof used (3.3.13). It is tempting to conjecture that a $k$ variable analogue of (3.3.13) holds in the $k$ variable case:

**Conjecture 3.6.2.** *For any $a_1 \dots a_k > 0$, there exist $p_1 \dots p_k > 0$ so that for all $x_1 \dots x_k > 0$,*

$$\Phi_k(x_1, \dots, x_k) \leqslant p_1 \hat{x}_1 + \cdots + p_k \hat{x}_k \tag{3.6.1}$$

*where $\hat{x}_i = \prod_{j \neq i} x_j$. In other words, the function*

$$\Theta_k(x_1, \dots, x_k) = x_1 \dots x_k \Phi_k(1/x_1, \dots, 1/x_k)$$

*is a concave function.*

Indeed, the author has verified Conjecture 3.6.2 for $\Phi_4$ and $\Phi_5$ with the aid of a computer. However, the full $k$-variable version remains unresolved. Fortunately, we can use a different argument to generalise (3.3.15).

**Lemma 3.6.3** (Generalising (3.3.15) to $k$ variables)**.** *Let $G$ be any finite abelian group, and let $f_1, \dots, f_k : G \to [0, 1]$ with $b_i = \sum_{g \in G} f_i(g)$. Then*

$$\begin{aligned} \sum_{\substack{g_1, \dots, g_k \in G \\ g_1 + \cdots + g_k = 0}} \Phi_k(f_1(g_1), \dots, f_k(g_k)) &\leqslant \Phi_k(b_1, \dots, b_k) \\ &= |G|^{k-1} \Phi(b_1/|G|, \dots, b_k/|G|) \end{aligned} \tag{3.6.2}$$

*Proof.* As in the proof of the case $k = 3$, the equality part of (3.6.2) follows from the homogeneity of $\Phi_k$. To prove the inequality, we will carry out an induction over $k$. For symmetrically decreasing functions $f, g : \mathbb{R} \to [0, 1]$, write $f \amalg g$ for the symmetrically decreasing function with the property that $\mu(\{x | f(x) \geqslant t\}) + \mu(\{x | g(x) \geqslant t\}) = \mu(\{x | (f \amalg g)(x) \geqslant t\})$, where $\mu(S)$ denotes the standard Lebesgue measure. In particular, $I_x \amalg I_y = I_{x+y}$.

For each $1 \leqslant i \leqslant k$, and each $g \in G$, let $f_i(g)$ be a symmetrically decreasing function, and let $h_i = \coprod_{g \in G} f_i(g)$. Then, it suffices to establish the following inequality:

$$\sum_{\substack{g_1, \dots, g_k \in G \\ g_1 + \cdots + g_k = 0}} (f_1(g_1) * \cdots * f_k(g_k))(0) \leqslant (h_1 * \cdots * h_k)(0). \tag{3.6.3}$$

Observe that for any particular value of $k$, (3.6.3) follows from (3.6.2) since we can express all of the symmetrically decreasing functions as combinations of intervals. In particular, the $k = 3$ case of (3.6.3) follows immediately from (3.3.15).

In order to establish (3.6.3) for a particular value of $k \geqslant 4$, define

$$F(g) = \sum_{\substack{g_3,\dots,g_k \in G \\ g_3+\dots+g_k=g}} (f_3(g_3) * \dots * f_k(g_k))(0).$$

Thus, the left hand side of (3.6.3) becomes

$$\sum_{\substack{g_1,g_2,g \in G \\ g_1+g_2+g=0}} (f_1(g_1) * f_2(g_2) * F(g))(0).$$

By the $k = 3$ case of (3.6.3), we immediately have

$$\sum_{\substack{g_1,g_2,g \in G \\ g_1+g_2+g=0}} (f_1(g_1) * f_2(g_2) * F(g))(0) \leqslant (h_1 * h_2 * H)(0),$$

where $H = \coprod_{g \in G} F(g)$. Thus, it will suffice to establish that

$$(h_1 * h_2 * H)(0) \leqslant (h_1 * h_2 * h_3 * \dots * h_k)(0)$$

or, by expressing $h_1 * h_2$ as a combination of intervals, that

$$(I_x * H)(0) \leqslant (I_x * h_3 * \dots * h_k)(0)$$

for any interval $I_x$.

But, by definition of $H$, it is possible to write $x = \sum_{g \in G} x_g$ in such a way that

$$\begin{aligned}(I_x * H)(0) &= \sum_{g \in G} (I_{x_g} * F(g))(0) \\ &= \sum_{\substack{g,g_3,\dots,g_k \in G \\ g=g_3+\dots+g_k}} (I_{x_g} * f_3(g_3) * \dots * f_k(g_k))(0) \\ &\leqslant (I_x * h_3 * \dots * h_k)(0),\end{aligned}$$

where the last line of this deduction follows from the induction hypothesis. This completes the proof of (3.6.3), and thus of Lemma 3.6.3. □

*Remark.* The definition of $f \amalg g$ in the proof of Lemma 3.6.3 is inspired by arguments of Hardy and Littlewood [32]. In particular, it is a continuous analogue of the following opperation for sequences: given two finite sequences $(a_i)$ and $(b_i)$ which are symmetrically decreasing (which in this setting means that $a_0 \geqslant a_1 \geqslant a_{-1} \geqslant a_2 \geqslant \dots$ and the same for the $b_i$), take the symmetrically decreasing sequence whose terms are the union of the terms in the sequences $(a_i)$ and $(b_i)$.

## 3.7 Systems of more than one equation

Another way in which one might wish to extend Theorem 3.1.4 is to ask if a similar result holds for systems of $m$ equations in $k$ variables. One might imagine that a result of the following form ought to hold.

*Question.* Suppose that $k \geqslant m+2$ and $m \geqslant 1$. Does there exist an explicit positive constant $\sigma_{m,k}$ with the following properties:

- For any system $\mathcal{A}$ of $m$ equations in $k$ variables, there are large sets $S$, for which there are at least $(\sigma_{m,k} - o(1))|S|^{k-m}$ $k$-tuples in $S$ satisfying $\mathcal{A}$.
- For any $\varepsilon > 0$, there are systems such that the number of $k$-tuples satisfying $\mathcal{A}$ in *any* large $S \subseteq \mathbb{Z}$ is no more than $(\sigma_{m,k} + \varepsilon)|S|^{k-m}$.

Thus, Theorems 3.1.4 and 3.5.4 tell us that $\sigma_{m,k}$ exists whenever $m = 1$, and that $\sigma_{1,k} = \sigma_k/k^{k-2}$ with $\sigma_k$ as defined by (3.5.2). However, it turns out that when $m > 1$, not even the first of these has a positive answer, in the following sense.

**Theorem 3.7.1.** *Let $\varepsilon > 0$. Then, there exists a non-degenerate system of two equations in four variables with the property that for any large enough $S$, there are no more than $\varepsilon|S|^2$ solutions to the system in $S$.*

*Remark.* An easy modification of the proof of Theorem 3.7.1 gives the analogous result for any choice of $k, m$ with $k \geqslant m+2$ and $m > 1$.

The goal of this section is to prove Theorem 3.7.1.

*Proof.* We will prove Theorem 3.7.1 for the following system:

$$\begin{aligned} x + y &= z \\ x + My &= w, \end{aligned} \tag{3.7.1}$$

where $M$ is a sufficiently large constant (in terms of $\varepsilon$) to be chosen later.

We will start by borrowing the following lemma, which appears as part of the proof of the Balog-Szemerédi-Gowers theorem.

**Lemma 3.7.2** ([54, Corollary 6.20])**.** *Let $G$ be a bipartite graph with vertex sets $A$ and $B$ and edge set $E \subseteq A \times B$. Suppose that $|E| \geqslant \varepsilon|A||B|$, for some $\varepsilon > 0$. Then we can find subsets $A' \subseteq A$ and $B' \subseteq B$, with $|A'| \gg_\varepsilon |A|$ and $|B'| \gg_\varepsilon |B|$, such that, whenever $a \in A'$ and $b \in B'$, there are $\gg_\varepsilon |A||B|$ paths of length three from $a$ to $b$ in $G$.*

Let $S$ be a sufficiently large set (in terms of $M$ and $\varepsilon$), and suppose that there are more than $\varepsilon|S|^2$ solutions to (3.7.1) in $S$. Consider the bipartite graph on vertex set $A \amalg B$, where $A = B = S$; that is, both parts of $G$ are $S$. Draw an edge from $a$ to $b$ if and only if there is a solution to (3.7.1) with $x = a$ and $y = b$; in other words, if $a + b$ and $a + Mb$ are both in $S$. In particular, $G$ has at least $\varepsilon|S|^2$ edges.

We may immediately apply Lemma 3.7.2 to $G$. This gives us sets $A' \subseteq A$ and $B' \subseteq B$ such that, for any $a \in A'$ and $b \in B'$, there are $\gg_\varepsilon |S|^2$ paths of length 3 in $G$ from $a$ to $b$.

**Claim.** *These sets $A'$ and $B'$ satisfy $|A' + B'| \ll_\varepsilon |S|$ and $|A' + M \cdot B'| \ll_\varepsilon |S|$.*

*Proof of Claim.* To prove this claim, we can use an argument similar to that used in the proof of the Balog-Szemerédi-Gowers theorem. Showing that $|A' + B'| \ll_\varepsilon |S|$ and $|A' + M \cdot B'| \ll_\varepsilon |S|$ are similar, so we will only do the former.

Let $X$ denote the set of triples $(x, y, z)$ of elements of $(A + B) \cap S$, for which $x - y + z \in A' + B'$. We may trivially upper bound $|X|$; indeed, $|(A + B) \cap S| \leqslant |S|$, so $|X| \leqslant |S|^3$.

For a lower bound on $|X|$, consider an element $a + b$ of $A' + B'$. By definition, there are $\gg_\varepsilon |S|^2$ paths of length 3 from $a$ to $b$ in $G$. Each such path may be written $a \sim b', a' \sim b', a' \sim b$ for some $a' \in A, b' \in B$. In other words, $a + b', a' + b'$ and $a + b'$ are all in $S$.

Now, $(a + b') - (a' + b') + (a' + b) = (a + b)$, so we have located a triple $x, y, z \in (A + B) \cap S$ with $x - y + z = a + b$. These triples will be different for different paths, and so there must be $\gg_\varepsilon |S|^2$ such triples.

There are $|A' + B'|$ elements of $A' + B'$, each of which gives $\gg_\varepsilon |S|^2$ triples $x, y, z$. Thus, we have that $|A' + B'||S|^2 \ll_\varepsilon |S|^3$, and thus $|A' + B'| \ll_\varepsilon |S|$, as required. $\Diamond$

Let us now see how we may use this claim to complete the proof of Theorem 3.7.1. Lemma 3.2.4 immediately tells us that $|B' - M \cdot B'| \ll_\varepsilon |S|$, and thus that $|B' - M \cdot B'| \ll_\varepsilon |B'|$. This contradicts Theorem 3.2.7, provided that $M$ is sufficiently large.

$\square$

## 3.8 The case $x + 2y + z = 0$

In this section, we will discuss how to complete the proof of (3.1.4), namely that $\gamma_{1,2,1} = \frac{1}{2}$, by using Lemma 3.1.5. Lemma 3.1.5 tells us that

$$\gamma_{1,2,1} = \sup_{n,f} \left( \sum_{\substack{x,y,z \in \mathbb{Z}/2^n\mathbb{Z} \\ x+2y+z=0}} \frac{1}{2}\Phi(f(x), 2f(y), f(z)) \right) \Big/ \left( \sum_x f(x) \right)^2 \tag{3.8.1}$$

Now, we will prove that $\gamma_{1,2,1} = \frac{1}{2}$ by showing that

$$\sum_{\substack{x,y,z \in \mathbb{Z}/2^n\mathbb{Z} \\ x+2y+z=0}} \frac{1}{2}\Phi(f(x), 2f(y), f(z)) \leqslant \sum_{x,y \in \mathbb{Z}/2^n\mathbb{Z}} \frac{1}{2}\Phi(f(x), 2f(y), f(x)). \tag{3.8.2}$$

It is easy to see why (3.8.2) would imply that $\gamma_{1,2,1} = \frac{1}{2}$. This is because the right hand side of (3.8.2) becomes

$$\begin{aligned} \sum_{x,y} \frac{1}{2}\Phi(f(x), 2f(y), f(x)) &= \frac{1}{2} \sum_{x,y} \frac{1}{2}\Big(\Phi(f(x), 2f(y), f(x)) + \Phi(f(y), 2f(x), f(y))\Big) \\ &= \frac{1}{2} \sum_{x,y} f(x)f(y) \\ &= \frac{1}{2} \left( \sum_x f(x) \right)^2, \end{aligned}$$

where the second equality follows from (3.3.11).

It remains to establish (3.8.2), and to do so we will use the following Lemma.

**Lemma 3.8.1.**

- *Let $a_0, a_1, b_0, b_1, c_0, c_1$ be positive real numbers with the property that $a_0 \leqslant a_1$, $b_0 \leqslant b_1$ and $c_0 \leqslant c_1$. Then $\Phi(a_i, b_j, c_k) + \Phi(a_{1-i}, b_{1-j}, c_{1-k})$ is maximized when $i = j = k$.*

- *Let $(a_i)_1^N$ and $(b_i)_1^N$ be two increasing sequences of $N$ positive real numbers, and let $c > 0$. Then, over all permutations $\tau$ of $[N]$, the quantity $\sum_i \Phi(a_i, b_{\tau(i)}, c)$ is maximized when $\tau$ is the identity permutation.*

*Proof.*

- To prove the first part, observe that without loss of generality, it suffices to show that

$$\Phi(a_0, b_0, c_1) + \Phi(a_1, b_1, c_0) \leqslant \Phi(a_0, b_0, c_0) + \Phi(a_1, b_1, c_1). \tag{3.8.3}$$

  Recall that by definition $\Phi(a, b, c) = I_a * I_b * I_c(0)$. Thus, we may rewrite (3.8.3) as follows:

$$I_{a_0} * I_{b_0} * I_{c_1}(0) + I_{a_1} * I_{b_1} * I_{c_0}(0) \leqslant I_{a_0} * I_{b_0} * I_{c_0}(0) + I_{a_1} * I_{b_1} * I_{c_1}(0).$$

  Subtracting the left hand side from the right hand side, we obtain

$$(I_{a_1} * I_{b_1} - I_{a_0} * I_{b_0}) * (I_{c_1} - I_{c_0})(0),$$

  which is nonnegative as both $(I_{a_1} * I_{b_1} - I_{a_0} * I_{b_0})$ and $(I_{c_1} - I_{c_0})$ are pointwise nonnegative.

- Choose the permutation $\tau$ so as to maximise $\sum_i \Phi(a_i, b_{\tau(i)}, c)$, and among all such choices of $\tau$, choose the one to maximise $\sum_i i\tau(i)$. If $\tau$ is not the identity, then there must exist $j < k$ such that $\tau(j) > \tau(k)$.

  Now, let $\tau' = \tau$, except with $\tau'(j) = \tau(k)$ and $\tau'(k) = \tau(j)$. Then $\sum i\tau(i) < \sum_i i\tau'(i)$, and thus the inequality $\sum_i \Phi(a_i, b_{\tau(i)}, c) \geqslant \sum_i \Phi(a_i, b_{\tau'(i)}, c)$ must be strict.

  It follows that

$$\Phi(a_i, b_{\tau(i)}, c) + \Phi(a_j, b_{\tau(j)}, c) > \Phi(a_i, b_{\tau'(i)}, c) + \Phi(a_j, b_{\tau'(j)}, c),$$

  contradicting the previous part.

□

Finally, we will complete the proof of (3.8.2). To do this is now simple. We may rewrite the left hand side of the equation as

$$\sum_{\substack{x,y,z \in \mathbb{Z}/2^n\mathbb{Z} \\ x+2y+z=0}} \frac{1}{2}\Phi(f(x), 2f(y), f(z)) = \frac{1}{2}\sum_y \left( \sum_x \Phi(f(x), 2f(y), f(\tau_y(x))) \right)$$

where $\tau_y$ is defined to be the permutation $x \mapsto 2y - x$.

Then, Lemma 3.8.1 exactly tells us that

$$\sum_x \Phi(f(x), 2f(y), f(\tau_y(x))) \leqslant \sum_x \Phi(f(x), 2f(y), f(x)),$$

which completes the proof that $\gamma_{1,2,1} = \frac{1}{2}$.

*Remark.* Observe that this proof applies equally to the other two equations $x+2y-z = 0$ and $x - 2y + z = 0$ to prove that $\gamma_{1,-2,1} = \gamma_{1,2,-1} = \frac{1}{2}$.

## 3.9 Equations of the form $x + y = kz$

The equation $x + y = kz$, where $k$ is a constant, is of particular special interest. Of course, Proposition 3.1.2 shows that for any $k$, we have $\gamma_{1,1,-k} \geqslant \frac{1}{12}$. However, it turns out that for this family, we can improve the value $\frac{1}{12}$.

**Proposition 3.9.1.** *For any value of $k$, we have $\gamma_{1,1,-k} \geqslant \frac{1}{5}$.*

*Proof.* To prove Proposition 3.9.1, we will modify the construction used in the proof of Proposition 3.1.2. The idea is that instead of taking three sets $S_1, S_2$ and $S_3$ as we did in Proposition 3.1.2, here we will take two sets $S_1$ and $S_2$, for which there are many solutions to $x + y = kz$ with $x$ and $y$ taken from $S_1$ and $z$ taken from $S_2$. Let $M$ be a large integer, which we assume to be divisible by 10. We will define

$$S_1 = k[-3M/10, 3M/10]$$
$$S_2 = [-M/5, M/5],$$

and define $S = S_1 \cup S_2$. Then, $|S|$ is no more than $M$.

Now, we may find a large collection of triples $(x, y, z)$ by choosing $x$ and $y$ in $S_1$ arbitrarily, and selecting those for which $z = \frac{x+y}{k}$ is in $S_2$. We may compute the number of such triples by using Proposition 3.5.3; this tells us that the number is $\Phi(3M/5, 3M/5, 2M/5) + O(M)$.

We may use (3.3.11) to show that $\Phi(3M/5, 3M/5, 2M/5) = M^2/5$, completing the proof of Proposition 3.9.1. □

*Remark.* A similar construction works if $k$ is a rational number that is not an integer; simply scale both $S_1$ and $S_2$ by the denominator of $k$.

In this section, we will prove an analogue of Theorem 3.1.4 for this family of equations.

**Theorem 3.9.2.** *The constant $\frac{1}{5}$ in the statement of Proposition 3.9.1 is optimal, in the following sense. For any $\varepsilon > 0$, there exists a choice of $k$ for which $\gamma_{1,1,-k} \leqslant \frac{1}{5} + \varepsilon$.*

*Proof.* We will prove Theorem 3.9.2 by modifying the elementary proof of Theorem 3.1.4. As in that argument, we start with a sufficiently large set $S$ and apply Theorem 3.2.2 to obtain a decomposition $S = S_1 \amalg \ldots \amalg S_n \amalg S_0$. Again, we can show that the contribution to $T(S)$ from solutions involving $S_0$ is small.

This allows us to draw the labelled digraph $G$ whose vertices correspond to the $S_i$ with $i \geqslant 1$, and an edge from $S_i$ to $S_j$ with label 1, $k^{-1}$ or $k$ if $T(S_i, S_j, S)$, $T(S_i, 1, S_j)$ or $T(S_j, 1, S_i)$ respectively is large enough (as in the proof of Theorem 3.1.4). An analogue of Lemma 3.2.9 shows that we can ignore solutions which do not come from triples $S_i$, $S_j$ and $S_k$ with the relevant edges present. Then, an analogue of Lemma 3.2.10 allows us to prove that the product of the labels along any cycle in $G$ must be 1, with the same proof, provided that $k$ is suitably large for the given $\varepsilon$.

At this point, our argument diverges from the argument to prove Theorem 3.1.4; for instance, it is no longer the case that $G$ can have no loops (as it is allowed to have edges with label 1). However, we can still define a function

$$d : [n] \to \mathbb{Q}$$

with the property that if $S_i \to S_j$ has label $t$, then $d(j) = td(i)$. We do this as we did in the proof of Lemma 3.2.11, and observe that for each $i \in [n]$, we have $d(i) = k^r$ for some integer $r$.

Now, let $R_r = \cup_{i:d(i)=k^r} S_i$, and so an upper bound for the total number of solutions under consideration is

$$\sum_r T(R_r, R_r, R_{r-1}),$$

where the sum is taken over all $r \in \mathbb{Z}$ with $R_r$ and $R_{r-1}$ nonempty.

Now, set $s_r = |R_r|$. In order to obtain an upper bound for $T(R_r, R_r, R_{r-1})$, we will use [43, Theorem 1]; recall that this says that we may assume that $R_r$ and $R_{r-1}$ are arithmetic progressions roughly centered at the origin with lengths $s_r$ and $s_{r-1}$. Combining this with Proposition 3.5.3, we deduce that

$$T(R_r, R_r, R_{r-1}) \leqslant \Phi(s_r, s_r, s_{r-1}) + O(|S|).$$

Summing over $r$, it follows that

$$\sum_r T(R_r, R_r, R_{r-1}) \leqslant \sum_r \Phi(s_r, s_r, s_{r-1}) + O_\varepsilon(|S|)$$

since $n$ is bounded depending only on $\varepsilon$. Thus, it suffices to prove that for any sequence of nonnegative real numbers $s_r$ with finite support, we have that

$$\sum_r \Phi(s_r, s_r, s_{r-1}) \leqslant \frac{1}{5}\left(\sum s_i\right)^2. \tag{3.9.1}$$

To prove that (3.9.1) holds, we start by proving the following lemma, which will enable us to reduce to the case of a single nonzero term on the left hand side.

**Lemma 3.9.3.** *Given any nonnegative real numbers $a$, $b$ and $c$, there are nonnegative real numbers $x$ and $y$ with $x+y = a+b+c$, $x \geqslant a$ and $\Phi(a,a,b)+\Phi(b,b,c) \leqslant \Phi(x,x,y)$.*

It is easy to see how Lemma 3.9.3 will enable us to reduce the left hand side (3.9.1) to a single term. Indeed, suppose that $r$ is the smallest index for which $s_r$ is nonzero. Apply the lemma with $a = s_{r+2}$, $b = s_{r+1}$ and $c = s_r$ to find $x$ and $y$, and replace $s_{r+2}$, $s_{r+1}$ and $s_r$ with $x$, $y$ and 0 respectively.

The right hand side of (3.9.1) remains unchanged since $x+y = a+b+c$. However, the left hand side does not decrease, as $\Phi(s_{r+3}, s_{r+3}, s_{r+2}) \leqslant \Phi(s_{r+3}, s_{r+3}, x)$ by the monotonicity of $\Phi$, and $\Phi(s_{r+2}, s_{r+2}, s_{r+1}) + \Phi(s_{r+1}, s_{r+1}, s_r) \leqslant \Phi(x,x,y)$. Thus, we have an instance of (3.9.1) with one fewer term, and can iterate to reduce to the case of one nonzero term.

*Proof of Lemma 3.9.3.* First, we will specialise the formula (3.3.11) to $\Phi(x,x,y)$:

$$\Phi(x,x,y) = \begin{cases} x^2 & y \geqslant 2x \\ xy - \frac{1}{4}y^2 & \text{otherwise.} \end{cases} \tag{3.9.2}$$

Now, we will split into cases according to the size of $a$. First, suppose that $a \leqslant \frac{3}{5}(a+b+c)$. Then, we may take $x = \frac{3}{5}(a+b+c)$ and $y = \frac{2}{5}(a+b+c)$, so $x+y = a+b+c$, $x \geqslant a$ and $\Phi(x,x,y) = \frac{1}{5}(a+b+c)^2$. We will prove that $\Phi(a,a,b) + \Phi(b,b,c) \leqslant \frac{1}{5}(a+b+c)^2$ for any triple $a, b$ and $c$.

If $c > 2b$, then we may reduce $c$ to equal $2b$ - this will not affect the left hand side and will decrease the right hand side. Thus, we can assume $\Phi(b,b,c) = bc - \frac{1}{4}c^2$. Now, if $b \leqslant 2a$ then $\Phi(a,a,b) = ab - \frac{1}{4}b^2$, and the assertion that $ab - \frac{1}{4}b^2 + bc - \frac{1}{4}c^2 \leqslant \frac{1}{5}(a+b+c)^2$ is a rearrangement of $(2a-3b+2c)^2 + 5c^2 \geqslant 0$. On the other hand, if $b \geqslant 2a$ then $\Phi(a,a,b) = a^2$, and the assertion that $a^2 + bc - \frac{1}{4}c^2 \leqslant \frac{1}{5}(a+b+c)^2$ is a rearrangement of $(3c-2b)^2 + 8ac + 8a(b-2a) \geqslant 0$.

Otherwise, we have that $a > \frac{3}{5}(a+b+c)$, in which case we may take $x = a$ and $y = b+c$. Then, we will certainly have $x+y = a+b+c$ and $x \geqslant a$. Moreover, we have $b \leqslant b+c < 2a$ and so $\Phi(a,a,b) = ab - \frac{1}{4}b^2$ and $\Phi(a,a,b+c) = a(b+c) - \frac{1}{4}(b+c)^2$, so we must prove that $ab - \frac{1}{4}b^2 + \Phi(b,b,c) \leqslant a(b+c) - \frac{1}{4}(b+c)^2$, or equivalently that $\Phi(b,b,c) \leqslant ac - \frac{1}{2}bc - \frac{1}{4}c^2$.

To do this, we substitute $a \geqslant \frac{3}{2}(b+c)$ so we are required to prove that $\Phi(b,b,c) \leqslant bc + \frac{5}{4}c^2$, and this holds because $\Phi(b,b,c) \leqslant bc$. □

Finally, we must prove that $\Phi(x,x,y) \leqslant \frac{1}{5}(x+y)^2$. To do this, observe that if $y \geqslant 2x$ then the left hand side is $x^2$ and the right hand side is at least $\frac{9}{5}x^2$. On

the other hand, if $y \leqslant 2x$ then $\Phi(x,x,y) = xy - \frac{1}{4}y^2$, and then the inequality is a rearrangement of the assertion that $\frac{1}{20}(2x-3y)^2 \geqslant 0$.

This completes the proof of Theorem 3.9.2. □

### 3.9.1 The set of numbers of the form $\gamma_{a_1,a_2,a_3}$

Let $X \subseteq \left(\frac{1}{12}, \frac{3}{4}\right]$ be the set of numbers of the form $\gamma_{a_1,a_2,a_3}$ for nonzero coprime integers $a_1, a_2$ and $a_3$. Examining the proofs of Proposition 3.9.1 and Theorem 3.9.2 allows us to prove further results about the structure of the $X$. In particular, we can prove the following lemma.

**Lemma 3.9.4.** *Let $\ell$ be a fixed rational number. There is a constant $\eta_\ell > \frac{1}{12}$ with the following properties.*

1. *Whenever $a_2 = \ell a_1$, we have that $\gamma_{a_1,a_2,a_3} > \eta_\ell$.*
2. *Let $\varepsilon > 0$. Suppose that $k$ is a rational number with height large enough depending on $\varepsilon$, and $a_3 = ka_1$. Then, we have that $\gamma_{a_1,a_2,a_3} < \eta_\ell + \varepsilon$.*

For example, it can be deduced from the proof of Theorem 3.9.2 that $\eta_1$ exists and is equal to $\frac{1}{5}$. Furthermore, the elementary proof of Theorem 3.1.4 shows that if the ratios between $a_1, a_2$ and $a_3$ all have large height, then $\gamma_{a_1,a_2,a_3}$ is close to $\frac{1}{12}$. Consequently, whenever $\ell$ has large height it follows that $\eta_\ell$ must be close to $\frac{1}{12}$.

For any $\delta > 0$, there can be only finitely many values of $\ell$ for which $\eta_\ell > \frac{1}{12} + \delta$. This allows us to deduce the following structural result regarding $X$.

**Proposition 3.9.5.** *Any subset of $X$ has a largest element. In other words, $X$ has no infinite increasing sequence.*

*Remark.* Proposition 3.9.5 may also be phrased as saying that the set $-X$ is well-ordered.

Proposition 3.9.5 is a simple consequence of Lemma 3.9.4 and the above remarks. Indeed, the second part of Lemma 3.9.4 says that there can be no infinite increasing sequence of $\gamma_{a_1,a_2,a_3}$ with $\frac{a_1}{a_2}$ fixed. Thus, given an infinite increasing sequence of elements $\gamma_{a_1^{(i)},a_2^{(i)},a_3^{(i)}}$ of $X$, the heights of the ratios between the $a_1^{(i)}, a_2^{(i)}$ and $a_3^{(i)}$ must all tend to infinity with $i$. It follows that $\gamma_{a_1^{(i)},a_2^{(i)},a_3^{(i)}}$ must tend to $\frac{1}{12}$, contradicting the assertion that the sequence is increasing.

We now sketch the proof of Lemma 3.9.4.

*Sketch proof of Lemma 3.9.4.* Fix a rational number $\ell$, and let $\varepsilon > 0$. Suppose that $k$ is a rational number with height large enough depending on $\varepsilon$, and let $a_1, a_2$ and $a_3$ be such that $a_2 = \ell a_1$ and $a_3 = ka_1$.

Let $S$ be a set with many solutions to $a_1x_1 + a_2x_2 + a_3x_3 = 0$. We mimic the start of the proof of Theorem 3.9.2 (which itself mimics the start of the elementary proof of Theorem 3.1.4). This gives us a labelled digraph $G$, whose vertices are the sets $(S_i)_{i=1}^n$ of a decomposition of $S$, and with edges controlling which terms $T(S_i, S_j, S_k)$ are large; the total contribution to $T(S)$ from the other triples is no greater than $\varepsilon|S|^2$.

It will not be possible to prove that the product of the labels of the edges of a cycle is 1. However, if $k$ is large enough, it will be the case that the product of the labels of the edges of a cycle must be of the form $\ell^i$, for some $-n \leqslant i \leqslant n$.

Again, this allows us to define a function $d : [n] \to \mathbb{Q}$ with the property that if $S_i \to S_j$ has label $t$, then $\frac{d(j)}{td(i)}$ is either $\ell^{-1}$, 1 or $\ell$. We do this again as we did in the proof of Lemma 3.2.11, and again observe that for each $i \in [n]$, we have $d(i) = k^r$ for some integer $r$.

Letting $R_r = \cup_{i:d(i)=k^r} S_i$, we deduce the same upper bound for the total number of solutions:

$$\begin{aligned}\sum_r T_{a_1,\ell a_1,ka_1}(R_r, R_r, R_{r-1}) &= \sum_r T_{a_1,\ell a_1,a_1}(R_r, R_r, k \cdot R_{r-1}) \\ &= \sum_r T_{a_1,\ell a_1,a_1}(R'_r, R'_r, R'_{r-1}),\end{aligned}$$

where we set $R'_r = k^{-r} \cdot R_r$. In particular, as the sequence of sets $(R'_r)$ runs over all sequences of sets of rational numbers with fixed total size $|S|$, the maximal value of the expression on the right hand side is independent of $k$. Let $f(|S|)$ denote that maximum value (which depends only on $\ell$ and not $a_1$ or $k$), and define $\eta_\ell$ by

$$\eta_\ell = \limsup_{|S|\to\infty} \frac{f(|S|)}{|S|^2}.$$

Then, we have proven that an upper bound for $\gamma_{a_1,a_2,a_3}$ is given by $\eta_\ell + \varepsilon$.

Finally, observe that $\eta_\ell$ is a lower bound for $\gamma_{a_1,a_2,a_3}$. Indeed, given a sequence of sets $(R'_r)$, the set $S = \cup_r k^r \cdot R'_r$ has $\frac{T(S)}{|S|^2} > \eta_\ell$. Note that, in view of the construction in Proposition 3.1.2, this shows that $\eta_\ell > \frac{1}{12}$ and indeed completes the proof of Lemma 3.9.4. □

# Chapter 4

# Cyclically covering subspaces in $\mathbb{F}_2^n$

## 4.1 Introduction

For $n \in \mathbb{N}$, let $\{e_0, e_1, \ldots, e_{n-1}\}$ be the standard basis for $\mathbb{F}_2^n$. Throughout this chapter, the indices of vectors in $\mathbb{F}_2^n$ will be taken modulo $n$ (in particular, we set $e_n = e_0$). Define the *cyclic shift* operator $\sigma : \mathbb{F}_2^n \to \mathbb{F}_2^n$ by

$$\sigma\left(\sum_{i=0}^{n-1} x_i e_i\right) = \sum_{i=0}^{n-1} x_i e_{i+1}.$$

We say that a subspace $U \leqslant \mathbb{F}_2^n$ is *cyclically covering* if $\bigcup_{i=0}^{n-1} \sigma^i(U) = \mathbb{F}_2^n$. For any $n \in \mathbb{N}$, let $h_2(n)$ denote the largest possible codimension of a cyclically covering subspace of $\mathbb{F}_2^n$.

For all odd $n \geqslant 5$, Cameron, Ellis and Raynaud [11] give an explicit construction for a cyclically covering subspace with codimension 2, which establishes the lower bound $h_2(n) \geqslant 2$. In 1991, Cameron asked (a question equivalent to) whether $h_2(n) \to \infty$ as $n \to \infty$ over the odd integers or whether $h_2(n) = 2$ for infinitely many odd $n$ (see [10, Problem 190]). The motivation for the original formulation came from proving lower bounds for Isbell's conjecture (stated in [12]), and we refer the reader to the paper by Cameron, Ellis and Raynaud [11] for further discussion.

Cameron, Ellis and Raynaud [11] show that $h_2(mn) \geqslant \max\{h_2(m), h_2(n)\}$ for $m, n \in \mathbb{N}$. It follows that there are three possibilities for the behaviour of $h_2(n)$ as $n \to \infty$ over the odd integers:

1. either there is a prime $p > 2$ such that $h_2(p^n)$ is bounded over all $n \in \mathbb{N}$,

2. or there is some $M \in \mathbb{N}$ such that there are infinitely many primes $p$ with $h_2(p) \leqslant M$,

3. or $h_2(n) \to \infty$ over odd integers $n$.

In Section 4.2, we improve the stated lower bound by showing that $h_2(mn) \geqslant h_2(m) + h_2(n)$ for all $m, n \in \mathbb{N}$, which rules out the first case. The main result of this chapter shows that, provided there are infinitely many primes with 2 as a primitive root, the correct case is the second one.

**Theorem 4.1.1.** *Suppose that $p$ is a prime for which 2 is a primitive root. Then $h_2(p) = 2$.*

Artin conjectured that there are infinitely many primes for which 2 is a primitive root; such primes are known as Artin primes. More generally, Artin's conjecture states that for any $n \in \mathbb{N}$ which is not a perfect square, there are infinitely many primes $p$ such that $n$ is a primitive root modulo $p$. Artin's conjecture is widely believed; in particular, it follows as a consequence of the generalised Riemann hypothesis, as shown by Hooley [35]. While there are no primes for which Artin's conjecture has been established, Heath-Brown [34] has shown that there can be no more than two primes $q$ for which Artin's conjecture fails.

Throughout this chapter, rather than considering cyclically covering subspaces, we will consider an equivalent problem in the orthogonal complement. We say that a vector $v \in \mathbb{F}_2^n$ *works* if for every $x \in \mathbb{F}_2^n$ there is a $k$ (which we may take to be in $\{0, \dots, n-1\}$) such that $v \cdot \sigma^k x = 0$. We say that the vectors $v^{(1)}, v^{(2)}, \cdots, v^{(m)}$ *work together* if for every $x \in \mathbb{F}_2^n$ there is a $k \in \{0, \dots, n-1\}$ such that

$$v^{(1)} \cdot \sigma^k x = v^{(2)} \cdot \sigma^k x = \cdots = v^{(m)} \cdot \sigma^k x = 0.$$

Suppose $U$ is a cyclically covering subspace of $\mathbb{F}_2^n$ and let $v^{(1)}, \dots, v^{(m)}$ be a basis for $U^\perp$, the orthogonal complement of $U$. Since $U$ is cyclically covering, for any $x \in \mathbb{F}_2^n$, there exists $u \in U$ and a $k$ such that $x = \sigma^k u$, so, by definition,

$$v^{(i)} \cdot \sigma^{n-k} x = v^{(i)} \cdot u = 0 \quad \forall i \in [m]$$

and the vectors $\{v^{(1)}, \dots, v^{(m)}\}$ work together.

Conversely, if we have a set $V = \{v^{(1)}, \dots, v^{(m)}\}$ of vectors which work together, then $U = \mathrm{Span}(V)^\perp$ is cyclically covering. Indeed, if $x \in \mathbb{F}_2^n$, then by definition there is a $k$ such that

$$v^{(1)} \cdot \sigma^k x = v^{(2)} \cdot \sigma^k x = \cdots = v^{(m)} \cdot \sigma^k x = 0.$$

Hence, $\sigma^k x \in U$, and $x \in \sigma^{n-k} U$, which shows $U$ is cyclically covering.

This means that $h_2(n)$ is the largest value $m \in \mathbb{Z}$ such that there exist vectors $v^{(1)}, \ldots, v^{(m)} \in \mathbb{F}_2^n$ which work together and are linearly independent.

In Section 4.2 we use this formulation to prove the lower bound $h_2(mn) \geqslant h_2(m) + h_2(n)$, and to give an upper bound of $h_2(2n) \leqslant 2h_2(n)$. In this formulation, there is a direct correspondence between vectors that work and cyclically covering subspaces with codimension 1, and we give a precise characterisation of these vectors in Theorem 4.2.5.

Sections 4.3, 4.4 and 4.5 are all devoted to the proof of our main result, Theorem 4.1.1. Suppose that we could find three linearly independent vectors that work together for $p$ a prime with 2 as a primitive root. In Section 4.3 we show that we can assume one of the vectors is $\mathbf{e} = (0, 1, \ldots, 1)$, which allows us to prove the other vectors must have various properties. We call a vector $v \in \mathbb{F}_2^n$ *symmetric* if it is of the form $(v_0, v_1, v_2, \ldots, v_2, v_1)$ (in other words, $v_i = v_j$ whenever $i \equiv -j \mod n$). Based on our experimental results, we make the following conjecture.

**Conjecture 4.1.2.** *Suppose that $n \in \mathbb{N}$ is an odd number that is not divisible by 7. Let $v \in \mathbb{F}_2^n$, and suppose that $v$ and $\mathbf{e}$ work together. Then $v$ is symmetric.*

We also prove that if $\mathbf{e}$, $v$ and $w$ work together and are linearly independent, neither $v$ nor $w$ can be symmetric. Thus, Conjecture 4.1.2 would imply Theorem 4.1.1.

In Section 4.4, we give an equivalent graph-theoretic formulation of our conjecture, and we use this to prove various structural properties satisfied by a vector $v$ which works together with $\mathbf{e}$. In particular, we obtain some partial progress towards our conjecture by eliminating vectors $v$ which have large "girth".

In Section 4.5, we consider vectors with small "girth", and we show that we cannot have two linearly independent non-symmetric vectors with small girth that work together with $\mathbf{e}$. This then completes the proof of Theorem 4.1.1.

Cameron, Ellis and Raynaud [11] consider the more general problem of computing $h_q(n)$, the maximal codimension of a cyclically covering subspaces in $\mathbb{F}_q^n$, for prime powers $q$. Amongst many other results, they show that the lower bound $h_2(mn) \geqslant \max\{h_2(m), h_2(n)\}$ mentioned earlier holds even when 2 is replaced by any prime power $q$. In Section 4.6 we show how our lower bound also holds in $\mathbb{F}_q^n$ and extend the upper bound given in Theorem 4.2.3 to this setting as well. Finally, we show that $h_q(p) = 0$ when $q$ is an odd prime and $p > q$ is a prime with $q$ as a primitive root.

## 4.2 Multiplicative bounds

Cameron, Ellis and Raynaud [11, Lemma 3] proved that for any $m, n \in \mathbb{N}$, we have $h_2(mn) \geqslant \max\{h_2(n), h_2(m)\}$. We offer the following improvement to this lower bound.

**Theorem 4.2.1.** *For any $m, n \in \mathbb{N}$,*

$$h_2(mn) \geqslant h_2(m) + h_2(n)$$

Theorem 4.2.1 immediately tells us that $h_2(n_i) \to \infty$ for many sequences $n_i$.

**Corollary 4.2.2.** *$h_2(m^n) \to \infty$ as $n \to \infty$ for any $m$ for which $h_2(m) > 0$.*

Indeed, Theorem 4.2.1 immediately tells us that $h_2(m^n) \geqslant n h_2(m)$. Note that the condition $h_2(m) > 0$ is not very restrictive: Cameron, Ellis and Raynaud [11] proved that $h_2(m) = 0$ if and only if $m$ is a power of 2. We will provide an alternative proof of this fact later in this section.

*Proof of Theorem 4.2.1.* Let $v^{(1)}, \dots, v^{(a)} \in \mathbb{F}_2^m$ be a set of $a = h_2(m)$ linearly independent vectors which work together for $m$, and $w^{(1)}, \dots, w^{(b)} \in \mathbb{F}_2^n$ a set of $b = h_2(n)$ linearly independent vectors which work together for $n$. We will now give a family of $a + b$ linearly independent vectors in $\mathbb{F}_2^{mn}$ which work together for $mn$.

Firstly, for each $i \in [a]$ and $j \in [b]$, define the vectors $\widetilde{v}^{(i)} \in \mathbb{F}_2^{mn}$ and $\widetilde{w}^{(j)} \in \mathbb{F}_2^{mn}$ as

$$\widetilde{v}^{(i)} = (\underbrace{v_0^{(i)}, \dots, v_0^{(i)}}_{n \text{ copies}}, \dots, \underbrace{v_{m-1}^{(i)}, \dots, v_{m-1}^{(i)}}_{n \text{ copies}}),$$
$$\widetilde{w}^{(i)} = (\underbrace{w^{(i)}, \dots, w^{(i)}}_{m \text{ copies}}).$$

We claim that the family of vectors $\widetilde{v}^{(1)}, \dots, \widetilde{v}^{(a)}, \widetilde{w}^{(1)}, \dots, \widetilde{w}^{(b)}$ all work together and are linearly independent, which will imply that $h_2(mn) \geqslant a + b$.

First, we will prove that the vectors work together; in other words, we will prove that for any $x \in \mathbb{F}_2^{mn}$, there is a $k \in \mathbb{Z}$ such that $\widetilde{v}^{(i)} \cdot \sigma^k x = \widetilde{w}^{(j)} \cdot \sigma^k x = 0$ for each $i \in [a]$ and $j \in [b]$.

Suppose $x \in \mathbb{F}_2^{mn}$. Let $y \in \mathbb{F}_2^n$ be the vector defined by

$$y = \left( \sum_{r=0}^{m-1} x_{rn}, \sum_{r=0}^{m-1} x_{rn+1}, \dots, \sum_{r=0}^{m-1} x_{rn+n-1} \right).$$

This means that $\widetilde{w}^{(j)} \cdot \sigma^k x = w^{(j)} \cdot \sigma^k y$ for all $k \in \mathbb{Z}$. Since the $w^{(j)}$ work together in $\mathbb{F}_2^n$, there is some choice of $k \in \mathbb{Z}$ such that $w^{(j)} \cdot \sigma^k y = 0$ for each $j$, which means that $\widetilde{w}^{(j)} \cdot \sigma^k x = 0$ for each $j$. We may assume that $k = 0$ by replacing $x$ with $\sigma^k x$.

Let $z \in \mathbb{F}_2^m$ be given by

$$z = (x_0 + \cdots + x_{n-1}, x_n + \cdots + x_{2n-1}, \ldots, x_{mn-n} + \cdots + x_{mn-1}).$$

Then

$$\widetilde{v}^{(i)} \cdot \sigma^{n\ell} x = v^{(i)} \cdot \sigma^\ell z \text{ for all } i \in [a] \text{ and } \ell \in \mathbb{Z}.$$

Since $v^{(1)}, \ldots, v^{(a)}$ work together for $m$, there is some choice of $\ell \in [m]$ such that $v^{(i)} \cdot \sigma^\ell z = 0$ for every $i \in [a]$. We conclude

$$\begin{aligned} \widetilde{v}^{(i)} \cdot \sigma^{\ell n} x = v^{(i)} \cdot \sigma^\ell z = 0 \quad \forall i \in [a], \\ \widetilde{w}^{(i)} \cdot \sigma^{\ell n} x = w^{(i)} \cdot y = 0 \quad \forall i \in [b]. \end{aligned}$$

This shows the vectors $\widetilde{v}^{(1)}, \ldots, \widetilde{v}^{(a)}, \widetilde{w}^{(1)}, \ldots, \widetilde{w}^{(b)}$ all work together and it only remains to show that the vectors are linearly independent in $\mathbb{F}_2^{mn}$.

First, observe that the vectors $\widetilde{v}^{(i)} \in \mathbb{F}_2^{mn}$ for $i \in [a]$ are linearly independent because the vectors $v^{(i)} \in \mathbb{F}_2^m$ are linearly independent. Similarly, the family $\widetilde{w}^{(j)}$ is linearly independent because the family $w^{(j)}$ is.

To show that the vectors $\widetilde{v}^{(1)}, \ldots, \widetilde{v}^{(a)}, \widetilde{w}^{(1)}, \ldots, \widetilde{w}^{(b)}$ are linearly independent, it suffices to show that any vector that is in the span of the $\widetilde{v}^{(i)}$ and also in the span of the $\widetilde{w}^{(j)}$ is the zero vector. Suppose that $u$ is such a vector.

By considering the standard basis vector $x = e_0$ in the definition of working together for $v^{(1)}, \ldots, v^{(a)}$, it follows that there must be an $r$ such that $v_r^{(i)} = 0$ for each $i$. In particular, $\widetilde{v}_{rn+s} = 0$ for each $s \in \{0, 1, \ldots, n-1\}$.

However, the fact that the $w^{(j)}$ are linearly independent means that the only linear combination of the $\widetilde{w}^{(j)}$ for which $w_{rn+s} = 0$ for each $s \in \{0, 1, \ldots, n-1\}$ is zero. Thus, $u = 0$, which means that the vectors $\widetilde{v}^{(1)}, \ldots, \widetilde{v}^{(a)}, \widetilde{w}^{(1)}, \ldots, \widetilde{w}^{(b)}$ are linearly independent, as required. □

We also obtain the following rough upper bound.

**Theorem 4.2.3.** *For any $n \geqslant 0$,*

$$h_2(2n) \leqslant 2h_2(n). \tag{4.2.1}$$

*Proof.* Suppose $\{v^{(1)}, \ldots, v^{(a)}\} \subseteq \mathbb{F}_2^{2n}$ is a collection of $a = h_2(2n)$ linearly independent vectors which work together for $2n$ and write $v^{(i)} = (u^{(i)} + w^{(i)}, w^{(i)})$ for $u^{(i)}, w^{(i)} \in \mathbb{F}_2^n$. The vectors $u^{(1)}, \ldots, u^{(a)} \in \mathbb{F}_2^n$ work together. To see why, note that for any $y \in \mathbb{F}_2^n$, there is a $k$ such that

$$u^{(i)} \cdot \sigma^k y = v^{(i)} \cdot \sigma^k (y, y) = 0$$

for every $i$.

Without loss of generality $u^{(1)}, \ldots, u^{(\ell)}$ is a maximal linearly independent subset of $\{u^{(1)}, \ldots, u^{(a)}\}$. Then for $j > \ell$ we can find $\lambda_1, \ldots, \lambda_\ell$ such that $u^{(j)} = \lambda_1 u^{(1)} + \cdots + \lambda_\ell u^{(\ell)}$. We may replace $v^{(j)}$ with $\lambda_1 v^{(1)} + \cdots + \lambda_\ell v^{(\ell)} + v^{(j)}$ without changing the span of the $v^{(i)}$ or the fact that they work together. This will give the vector

$$\left(\lambda_1 w^{(1)} + \cdots + \lambda_\ell w^{(\ell)} + w^{(j)}, \lambda_1 w^{(1)} + \cdots + \lambda_\ell w^{(\ell)} + w^{(j)}\right).$$

Doing this for all $j > \ell$, we can assume that $u^{(1)}, \ldots, u^{(\ell)}$ are linearly independent (hence $\ell \leqslant h_2(n)$) and that $u^{(i)} = 0$ for $i > \ell$.

In particular, for each $i > \ell$, we have $v^{(i)} = (w^{(i)}, w^{(i)})$. We claim that the vectors $w^{(i)}$ with $i > \ell$ are linearly independent and work together. This implies that $h_2(2n) - h_2(n) \leqslant h_2(n)$, which gives the desired result.

The linear independence of the $w^{(i)}$ follows from the linear independence of the $v^{(i)}$. To show that the $w^{(i)}$ work together we must prove that, for any $y \in \mathbb{F}_2^n$, there is some $k$ such that $w^{(i)} \cdot \sigma^k y = 0$ for each $i > \ell$.

Since the $v^{(i)}$ work together, there must exist a $k < 2n$ such that $v^{(i)} \cdot \sigma^k (0, y) = 0$ for each $i > \ell$. But we also have that $v^{(i)} = (w^{(i)}, w^{(i)})$, so

$$v^{(i)} \cdot \sigma^k y = (w^{(i)}, w^{(i)}) \cdot \sigma^k (0, y) = w^{(i)} \cdot \sigma^k y.$$

Thus, for this choice of $k$, we have that $w^{(i)} \cdot \sigma^k y = 0$ for each $i > \ell$, as required. □

It is unlikely that (4.2.1) is tight in general; indeed, if $n$ is chosen such that $h_2(n) > \frac{1}{2}\log_2(n) + 1$, then the tightness of (4.2.1) would contradict the upper bound $h_2(2n) \leqslant \log_2(n) + 1$. However, we may combine Theorem 4.2.3 with the fact that $h_2(1) = 0$ to deduce the following.

**Corollary 4.2.4.** *$h_2(2^i) = 0$ for every $i \in \mathbb{N}$.*

Let $W(n)$ be the set of $v \in \mathbb{F}_2^n$ which work. The fact that $h_2(2^i) = 0$ implies that no nonzero vector can work whenever $n$ is a power of 2 (the zero vector always works). We determine the exact structure of $W(n)$ whenever $n$ is not a power of 2, which in particular implies $h_2(n) \geqslant 1$ when $n$ is not a power of 2.

**Theorem 4.2.5.** *For $n$ odd, the vectors that work are given by*

$$W(n) = \{v \in \mathbb{F}_2^n : |v| \equiv 0 \mod 2\}.$$

*For $a$ odd and $b \in \mathbb{N}$,*

$$W(a2^b) = \left\{ v_0^{(1)} \dots v_0^{(2^b)} v_1^{(1)} \dots v_1^{(2^b)} \dots v_{a-1}^{(1)} \dots v_{a-1}^{(2^b)} : v^{(1)}, \dots, v^{(2^b)} \in W(a) \right\}.$$

*Remark.* Since $W(1) = \emptyset$, this shows $W(2^b) = \emptyset$ for all $b \geqslant 0$, which agrees with Corollary 4.2.4.

*Proof.* Suppose first that $v \in \mathbb{F}_2^n$ (where $n$ is odd) is a vector that works. Setting $x = (1, \dots, 1)$, there must be some $k$ such that $v \cdot \sigma^k x = 0$. However, for every shift $k \in [n]$,

$$v \cdot \sigma^k x = |v|.$$

Hence any vector which works must satisfy $|v| \equiv 0 \mod 2$.

Conversely, suppose that $|v| \equiv 0 \mod 2$. Given $x \in \mathbb{F}_2^n$, the number of shifts $k$ with $v \cdot \sigma^k x = 1$ is given by

$$|\{k : v \cdot \sigma^k x \equiv 1 \mod 2\}| \equiv v \cdot \sum_{k=0}^{n-1} \sigma^k x \equiv |v||x| \mod 2.$$

Since $|v| \equiv 0 \mod 2$, this shows the number of shifts $k$ with $v \cdot \sigma^k x = 1$ is even. Since $n$ is odd, we find that there is an odd number of shifts $k$ with $v \cdot \sigma^k x = 0$. In particular, there is at least one such $k$, proving the first claim.

Suppose now that $n$ is of the form $a2^b$ for $a$ odd. Let $v \in W(n)$ be given and split it into components $v^{(1)}, \dots, v^{(2^b)}$ by writing

$$v = (v_0^{(1)}, \dots, v_0^{(2^b)}, v_1^{(1)}, \dots, v_1^{(2^b)}, \dots, v_{a-1}^{(1)}, \dots, v_{a-1}^{(2^b)}). \tag{4.2.2}$$

Let $w \in \mathbb{F}_2^{2^b}$ be the vector with elements $w_j = |v^{(j)}|$. Suppose that $w \neq 0$. By Corollary 4.2.4, we then know that there is some vector $z \in \mathbb{F}_2^{2^b}$ for which all shifts fail, or in other words that

$$w \cdot \sigma^k z = 1 \ \ \forall k \in [2^b].$$

Now consider the vector $x \in \mathbb{F}_2^n$ with $x_{j2^i} = x_{j2^i+1} = \dots = x_{(j+1)2^i-1} = z_j$ for $j \in \{0, \dots, a-1\}$. Then

$$\sigma^k x \cdot v = \sigma^k z \cdot w = 1 \ \ \forall k \in [n].$$

This contradicts the assumption that $v$ works. Hence we find that $w_j = |v^{(j)}| \equiv 0$ for all $j \in [2^b]$, which shows that $v^{(1)}, \ldots, v^{(2^b)} \in W(a)$.

Conversely, suppose $v$ is given in the form (4.2.2) for $v^{(1)}, \ldots, v^{(2^b)} \in W(a)$. Let $x \in \mathbb{F}_2^n$ be arbitrary, and split it into components $x^{(1)}, \ldots, x^{(2^b)}$ in the same way. We will show that

$$|\{k \in [a] : v \cdot \sigma^{2^b k} x = 0\}|$$

is odd. Note that

$$\begin{aligned}\sum_{k=1}^{a} v \cdot \sigma^{2^b k} x &= \sum_{k=1}^{a} \left( \sum_{j=1}^{a} v^{(j)} \cdot \sigma^k x^{(j)} \right) \\ &= \sum_{j=1}^{a} \left( \sum_{k=1}^{a} v^{(j)} \cdot \sigma^k x^{(j)} \right) \\ &\equiv \sum_{j=1}^{a} |v^{(j)}||x^{(j)}| \equiv 0.\end{aligned}$$

This shows that the number of $k \in [a]$ for which $v \cdot \sigma^{2^b k} x = 1$ has to be even. Since $a$ is odd, this proves our claim. □

## 4.3 Primitive roots and a structural conjecture

In order to prove Theorem 4.1.1, we must first think about the structure of $\mathbb{F}_2^n$. For this, it will be easiest to think about vectors as polynomials, as in [11].

Given a vector $v \in \mathbb{F}_2^n$, let its corresponding polynomial be

$$f_v(x) = v_0 + v_1 X + v_2 X^2 + \cdots + v_{n-1} X^{n-1} \in \mathbb{F}_2[X]/\langle X^n - 1 \rangle.$$

We say a set of polynomials $f_{v^{(1)}}, \ldots, f_{v^{(m)}}$ *work together* if and only if, for every vector $x \in \mathbb{F}_2^n$, there is a $k$ such that the coefficient of $X^k$ in $f_{v^{(i)}} f_x$ is 0 for each $i \in [m]$. The following proposition shows that the two definitions of working together are equivalent, and so we will use them interchangeably.

**Proposition 4.3.1.** *The vectors $v^{(1)}, \ldots, v^{(m)}$ work together if and only if the corresponding polynomials $f_{v^{(1)}}, \ldots, f_{v^{(m)}}$ work together.*

*Proof.* By definition, $v^{(1)}, \ldots, v^{(m)}$ work together if and only if, for every $x \in \mathbb{F}_2^n$, there is a $k$ such that $v^{(i)} \cdot \sigma^k x = 0$ for each $i \in [m]$.

Let $x'$ be the reverse of $x$ (so $x'_j = x_{-j}$ for each $j$). The constant coefficient of

$$\left(v_0 + v_1 X + \cdots + v_{n-1} X^{n-1}\right)\left(x_0 + x_1 X + \cdots + x_{n-1} X^{n-1}\right) \mod (X^n - 1)$$

is given by

$$v_0x_0 + v_1x_{n-1} + v_2x_{n-2} + \cdots + v_{n-1}x_1 = v \cdot x'.$$

Similarly,

$$v^{(i)} \cdot \sigma^k x = \sum_{j=0}^{n-1} v_j^{(i)} x_{j-k} = \sum_{j=0}^{n-1} v_j^{(i)} x'_{k-j}$$

is the coefficient of $X^k$ in $f_{v^{(i)}} f_x$.

Thus, $v^{(i)} \cdot \sigma^k x = 0$ for every $i \in [m]$ if and only if the coefficient of $X^k$ of $f_{v^{(i)}} f_{x'}$ is 0 for every $i \in [m]$. □

Since $\alpha f_v + f_w = f_{\alpha v + w}$ for all $\alpha \in \mathbb{F}_2$ and $v, w \in \mathbb{F}_2^n$, a collection of vectors $v^{(1)}, \ldots, v^{(m)}$ is linearly independent if and only if the corresponding polynomials $f_{v^{(1)}}, \ldots, f_{v^{(m)}}$ are linearly independent.

*Remark.* The $m = 1$ case of Proposition 4.3.1 gives an alternative approach to prove Theorem 4.2.5. Indeed, Proposition 4.3.1 tells us that the vector $v$ will work if and only if $f_v$ works. By definition, $f_v$ works if and only if, for every $x \in \mathbb{F}_2^n$, there is a $k$ with $k \in \{0, \ldots, n-1\}$ such that the $X^k$ coefficient of $f_v f_x$ is 0; in other words, $f_v$ works if and only if $f_v$ does not divide $1 + X + \cdots + X^{n-1}$ in $\mathbb{F}_2[X]/(X^n - 1)$.

Since $(1 + X)(1 + X + \cdots + X^{n-1}) = X^n + 1$, it can be shown that that the polynomial $f_v$ does not divide $1 + X + \cdots + X^{n-1}$ exactly when $(1 + X)^r$ divides $f_v$, where $(1 + X)^r$ is the highest power of $1 + X$ which divides $X^n + 1$.

Now, $r = 2^b$, where $2^b$ is the highest power of 2 which divides $n$. Thus, $f_v$ does not divide $1 + X + \cdots + X^{n-1}$ exactly when $(1 + X)^{2^b}$ divides $f_v$, which is equivalent to the description given in Theorem 4.2.5.

This alternative formulation in terms of polynomials is particularly useful when $n = p$ is a prime with 2 as a primitive root. This is because of the following well-known results.

**Lemma 4.3.2** ([45, Theorem 2.47 (ii)])**.** *Let $p$ be an odd prime and let* $\mathrm{ord}_p(2)$ *be the multiplicative order of* 2 mod $p$*. Then the polynomial $1 + X + \cdots + X^{p-1}$ splits as a product of $t$ distinct irreducible factors in $\mathbb{F}_2[X]$, where $t = \frac{p-1}{\mathrm{ord}_p(2)}$.*

*In particular, if* 2 *is a primitive root mod $p$, then $1 + X + \cdots + X^{p-1}$ is irreducible in $\mathbb{F}_2[X]$.*

**Lemma 4.3.3** ([45, Theorem 1.61])**.** *The quotient ring $\mathbb{F}_2[X]/(f)$ is a field if and only if the polynomial $f$ is irreducible over $\mathbb{F}_2$.*

The following lemma is key to our approach to proving Theorem 4.1.1, and is the only part of the proof which requires 2 to be a primitive root modulo $p$.

**Lemma 4.3.4.** *Let $p$ be a prime with 2 as a primitive root, and define the identity vector $\mathbf{e}$ to be the vector such that $\mathbf{e}_0 = 0$ and $\mathbf{e}_i = 1$ for $i \neq 0$. Suppose that the vectors $v^{(1)}, \ldots, v^{(m)}$ work together and are linearly independent. Then there is a collection of $m$ linearly independent vectors which work together, and which contains $\mathbf{e}$.*

*Proof.* Each polynomial $f_{v^{(i)}}$ works individually. By Theorem 4.2.5, this means $|v^{(i)}| \equiv 0 \bmod 2$, which means that $f_{v^{(i)}}$ is divisible by $1 + X$. Since none of the polynomials can be the zero polynomial, they must all be equivalent to a nonzero polynomial modulo $1 + X + \cdots + X^{p-1}$.

By Lemma 4.3.2, $X^p - 1$ factors into irreducibles as

$$X^p - 1 = (X + 1)(1 + X + \cdots + X^{p-1}).$$

Thus, $\mathbb{F}_2[X]/(X^p - 1) \cong \mathbb{F}_2[X]/(X + 1) \oplus \mathbb{F}_2[X]/(1 + X + \cdots + X^{p-1})$, and Lemma 4.3.3 tells us that the two factors are fields. In particular, the irreducibility of $1 + X + \cdots + X^{p-1}$ tells us that each $f_{v^{(i)}}$ is invertible modulo $1 + X + \cdots + X^{p-1}$.

Thus, we may use the Chinese Remainder Theorem to find a polynomial $q$ which is 1 mod $(1 + X)$ and an inverse to $f_{v^{(1)}}$ modulo $(1 + X + \cdots + X^{p-1})$.

Now, $qf_{v^{(1)}}$ is 0 mod $(X + 1)$ and 1 mod $(1 + X + \cdots + X^{p-1})$, so

$$qf_{v^{(1)}} = X + \cdots + X^{p-1} = f_{\mathbf{e}}.$$

We claim that the polynomials $qf_{v^{(1)}}, \ldots, qf_{v^{(m)}}$ still work together and are linearly independent. Taking the vectors in $\mathbb{F}_2^p$ corresponding to the given polynomials then gives the collection of vectors required by the statement of the lemma.

Suppose first that $qf_{v^{(1)}}, \ldots, qf_{v^{(m)}}$ do not work together; thus, there exists a vector $x \in \mathbb{F}_2^p$ such that for any $k$, there is an $i \in [m]$ such that the $X^k$ coefficient of $f_x q f_{v^{(i)}}$ is 1. Writing $f_x q = f_{x'}$ for some $x' \in \mathbb{F}_2^p$, we find that for any $k$, there is an $i \in [m]$ such that the $X^k$ coefficient of $f_{x'} f_{v^{(i)}}$ is 1. This contradicts the assumption that the $v^{(i)}$ work together.

Next, suppose that the $qf_{v^{(i)}}$ are not linearly independent. Then there is a relation

$$\sum_i \lambda_i q f_{v^{(i)}} \equiv 0 \bmod (X^p - 1)$$

for some choice of $\lambda_i \in \mathbb{F}_2$ not all 0. However, $q$ is invertible modulo $X^p - 1$ by construction, so

$$\sum_i \lambda_i f_{v^{(i)}} \equiv 0 \bmod (X^p - 1).$$

This contradicts the linear independence of the $f_{v^{(i)}}$. □

Lemma 4.3.4 tells us that in order to show that $h_2(p) = 2$ for a prime $p$ with 2 as a primitive root, it suffices to show that there is no collection of three linearly independent vectors $\mathbf{e}, v, w$ which work together.

We now turn to examining the structure of the set of vectors $v \in \mathbb{F}_2^n$ which work with $\mathbf{e}$. There is a fairly substantial collection of vectors $v$ which work with $\mathbf{e}$, as shown by the following proposition.

**Proposition 4.3.5.** *Let $n$ be any odd number, and suppose $v \in \mathbb{F}_2^n$ is a vector with $v_0 = 0$. Suppose further that $v$ is* symmetric*, in that $v_i = v_{-i}$ for each $i$. Then $v$ works together with $\mathbf{e}$.*

*Proof.* Let $x \in \mathbb{F}_2^n$ be any vector. We must prove that there exists a $k$ such that $\mathbf{e} \cdot \sigma^k x = v \cdot \sigma^k x = 0$.

First, observe that we may assume that $x$ has odd weight. If $x$ had even weight, we could add the vector $z = (1, \ldots, 1)$ of all ones; this would give a vector of odd weight but, since $v \cdot \sigma^k z = \mathbf{e} \cdot \sigma^k z = 0$, it does not change the inner products $v \cdot \sigma^k x$ and $\mathbf{e} \cdot \sigma^k x$.

Now, $\mathbf{e} \cdot \sigma^k x = 0$ exactly when $x_{-k} = 1$, and for such a choice of $k$,

$$v \cdot \sigma^k x = \sum_{i : x_{i-k} = 1} v_i. \tag{4.3.1}$$

Summing (4.3.1) over all of the values of $k$ with $x_{-k} = 1$, we obtain

$$\begin{aligned} \sum_{k : x_{-k} = 1} v \cdot \sigma^k x &= \sum_{(k,i) : x_{-k} = x_{i-k} = 1} v_i \\ &= \sum_{(j_1, j_2) : x_{j_1} = x_{j_2} = 1} v_{j_2 - j_1}. \end{aligned}$$

Since $v_0 = 0$, and $v_{j_2 - j_1} + v_{j_1 - j_2} = 0$, this sum is 0.

Therefore, the sum of $v \cdot \sigma^k x$ over all $k$ such that $\mathbf{e} \cdot \sigma^k x = 0$ is zero. Since there are an odd number of such $k$ (as $x$ has odd weight), at least one of these $k$ must have $\mathbf{e} \cdot \sigma^k x = v \cdot \sigma^k x = 0$. □

Since $\mathbf{e}$ and $v$ work together whenever $v$ is symmetric, the following lemma will be necessary if we are to prove that $h_2(p) = 2$.

**Lemma 4.3.6.** *Let $p > 3$ be a prime, and suppose that $v \neq \mathbf{e}$ is a nonzero symmetric vector. Suppose further that $w \in \mathbb{F}_2^n$ is any vector such that $\mathbf{e}, v$ and $w$ work together. Then $w$ is contained in the subspace $\langle \mathbf{e}, v \rangle$ spanned by $\mathbf{e}$ and $v$.*

In order to prove this, we will rely on the Cauchy-Davenport inequality [54, Theorem 5.4] and Vosper's theorem [54, Theorem 5.9], both of which we state here for convenience.

**Theorem 4.3.7** (Cauchy-Davenport Inequality)**.** *Suppose that $p$ is a prime, and $A$ and $B$ are two sets in $\mathbb{Z}/p\mathbb{Z}$. Then*

$$|A + B| \geqslant \min(|A| + |B| - 1, p).$$

**Theorem 4.3.8** (Vosper's Theorem)**.** *Suppose that $p$ is a prime, and $A$ and $B$ are two sets in $\mathbb{Z}/p\mathbb{Z}$. Suppose that $|A|, |B| \geqslant 2$ and $|A + B| \leqslant p - 2$. Then $|A + B| = |A| + |B| - 1$ if and only if $A$ and $B$ are arithmetic progressions with the same common difference.*

We will often implicitly use the following observation, which gives a simple restriction on vectors $v$ that can work together with $\mathbf{e}$.

**Observation 4.3.9.** *Suppose that $v \in \mathbb{F}_2^n$ works with $\mathbf{e}$. Then $v_0 = 0$.*

This follows by considering the test vector $x = e_0$ of weight 1. Since $\sigma^k x \cdot \mathbf{e}$ is only zero if $k = 0$ and $\mathbf{e}$ works with $v$, we must have that $v_0 = e_0 \cdot v = 0$.

*Proof of Lemma 4.3.6.* Suppose that $v \in \mathbb{F}_2^p$ is symmetric, nonzero and not equal to $\mathbf{e}$. Let $A \subseteq \mathbb{Z}/p\mathbb{Z}$ consist of the values of $i$ for which $v_i = 1$, and let $B \subseteq (\mathbb{Z}/p\mathbb{Z}) \setminus \{0\}$ consist of the values of $i$ (except 0) for which $v_i = 0$. Observe that both $A = -A$ and $B = -B$ since $v$ is symmetric, and that $|A|,\ |B| \geqslant 2$ as we assume that $v$ is nonzero and $v \neq \mathbf{e}$.

The space $\langle \mathbf{e}, v \rangle$ consists of the four vectors $u$ which satisfy both that $u_0 = 0$ and that $u$ is constant on both $A$ and $B$. Thus, if $w$ is a vector that works together with both $\mathbf{e}$ and $v$, it suffices to prove that $w$ is constant on both $A$ and $B$, since $w_0 = 0$ by Observation 4.3.9.

Consider the equivalence relation $\sim$ on $(\mathbb{Z}/p\mathbb{Z}) \setminus \{0\}$ generated as follows. We say that $i \sim j$ if $w_i = w_j$ for any $w$ which works together with $\mathbf{e}$ and $v$. Thus, it suffices to prove that the only two equivalence classes are $A$ and $B$. The following claim will help us to do so.

**Claim.** *Suppose that $i \in B$ and $j \in A$ are such that $i + j \in A$. Then $j \sim i + j$. The same also holds with $A$ and $B$ reversed.*

*Proof of Claim.* We only prove the first case as the second is almost identical. Consider the vector $x = e_0 + e_i + e_{-j}$ of weight 3. The only shifts of $x$ which are orthogonal to $\mathbf{e}$ are $x$, $\sigma^{-i}x$ and $\sigma^j x$, and by assumption neither $x$ nor $\sigma^{-i}x$ are orthogonal to $v$.

Thus the only shift of $x$ which is orthogonal to both $\mathbf{e}$ and $v$ is $\sigma^j x$, and this must therefore be orthogonal to $w$ for any vector $w$ that works together with $\mathbf{e}$ and $v$. Hence $w_j = w_{i+j}$, which means that $j \sim i + j$. $\lozenge$

Now, we prove that the only two equivalence classes are $A$ and $B$. We will prove that $A$ is an equivalence class; the proof that $B$ is an equivalence class is almost identical. Suppose that $C \subsetneq A$ is an equivalence class of $\sim$ which is not the whole of $A$; without loss of generality, we may assume that $C$ is the smallest such class, which means that $|C| \leqslant |A|/2$.

Suppose that $i \in B$ and $j \in C$. Then $i + j \neq 0$, so either $j + i \in B$, or $j + i \in A$. In the latter case, $j \sim j + i$, and so the claim tells us that $j + i \in C$. In particular, we always have $j + i \notin A \setminus C$. Thus, $B + C \subseteq B \amalg C$.

Now, letting $B_0 = B \amalg \{0\}$ and $C_0 = C \amalg \{0\}$, we see that

$$B_0 + C_0 \subseteq B \amalg C \amalg \{0\}. \tag{4.3.2}$$

Since $C \neq A$, it must be the case that $|B \amalg C \amalg \{0\}| < p$. Thus, by Cauchy-Davenport we have that $|B_0 + C_0| \geqslant |B_0| + |C_0| - 1 = |B| + |C| + 1$, and so we have equality in (4.3.2). Both of $B_0$ and $C_0$ have at least two elements, so we have two cases.

**Case 1:** $|B_0 + C_0| \geqslant p - 1$. Then $|B_0| + |C_0| = p$, and $A \setminus C$ has only one element. By assumption, $|C| \leqslant |A|/2$, and so we must have $|A| = 2$ and $|C| = 1$.

Since $A$ is symmetric, it is of the form $\{-a, a\}$ with either $a$ or $-a$ the unique element of $C$; without loss of generality, we can assume that $C = \{1\}$ and $A = \{-1, 1\}$. Since $p > 3$, we must have $2 \in B$, and applying the Claim with $i = 2$ and $j = -1$ gives $-1 \sim 1$. This contradicts the fact that $C$ is an equivalence class.

**Case 2:** $|B_0 + C_0| \leqslant p - 2$, and so we can apply Vosper's Theorem. This tells us that $B_0$ and $C_0$ are arithmetic progressions with the same common difference, which we may assume is 1 by scaling. Since 0 is part of the arithmetic progression $B_0$ and $B_0 = -B_0$, both 1 and $-1$ must be in $B_0$. However, $C_0$ is an arithmetic progression with common difference 1, so either 1 or $-1$ must be in $C_0$. This contradicts the fact that $C \subseteq A$ and $A$ and $B$ are disjoint.

We conclude that the only two equivalence classes must be $A$ and $B$. Therefore, any vector $w$ which works with both $\mathbf{e}$ and $v$ must be constant on both $A$ and $B$, which means that $w$ is contained in the span of $\mathbf{e}$ and $v$, as required. □

As well as using that $n$ is odd, the proof of Proposition 4.3.5 crucially relies on the fact that $v$ is symmetric. In view of this, it is tempting to conjecture that a converse is true; in other words, that the only vectors $v \in \mathbb{F}_2^n$ which work together with $\mathbf{e}$ are the symmetric vectors. This leads us to propose Conjecture 4.1.2. Indeed, if Conjecture 4.1.2 were true, then Lemma 4.3.4 and Lemma 4.3.6 would tell us immediately that $h_2(p) = 2$ for any prime $p$ with 2 as a primitive root.

Note that Conjecture 4.1.2 cannot hold for $p = 7$, as demonstrated by the following counterexample.

**Observation 4.3.10.** *In $\mathbb{F}_2^7$, the following two vectors work together:*

$$\mathbf{e} = (0, 1, 1, 1, 1, 1, 1)$$
$$v = (0, 1, 1, 0, 0, 0, 0).$$

We have verified Conjecture 4.1.2 for $3 \leqslant n \leqslant 43$, and have some partial progress towards it for arbitrary primes. We will present our partial progress in the next section, and then show how we can use it to prove Theorem 4.1.1 in Section 4.5.

## 4.4 Partial progress towards conjecture

In this section, we will show that any vector $v$ which is a counterexample to Conjecture 4.1.2 must have certain properties.

### 4.4.1 A further reformulation of the problem

First, we will provide a reinterpretation of the statement that $v$ works together with $\mathbf{e}$, which will help us to describe the structure of such a vector.

To any vector $v \in \mathbb{F}_2^n$ we associate a Cayley (multi)digraph $G_v$ on vertex set $\mathbb{Z}/n\mathbb{Z}$, where we draw a directed edge from $i$ to $i + j$ exactly when $v_j = 1$. By Observation 4.3.9, we may restrict to vectors $v \in \mathbb{F}_2^n$ with $v_0 = 0$, which means that $G_v$ will have no self-loops. The relevance of $G_v$ to our problem is given by the following proposition.

**Proposition 4.4.1.** *Let $n$ be odd. A vector $v \in \mathbb{F}_2^n$ with $v_0 = 0$ works together with $\mathbf{e}$ if and only if $G_v$ has no induced subgraph on an odd number of vertices where each vertex has odd outdegree.*

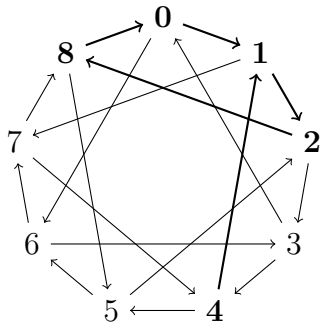



Figure 4.1: The Cayley graph $G_v$ for $v = (0, 1, 0, 0, 0, 0, 1, 0, 0)$. The graph induced on $\{0, 1, 2, 4, 8\}$ is of odd size and every vertex has odd outdegree so the vector $(1, 1, 1, 0, 1, 0, 0, 0, 1)$ shows that $\mathbf{e}$ and $v$ do not work together.

*Proof.* Suppose first that $v$ and $\mathbf{e}$ do not work together. Equivalently, there exists a vector $x \in \mathbb{F}_2^n$, which we may assume has odd weight, such that any shift $\sigma^k x$ is non-orthogonal to at least one of $\mathbf{e}$ and $v$.

Let $A = \{i : x_i = 1\}$, and consider $G_v[A]$. We have that $\sigma^{-k}x$ is orthogonal to $\mathbf{e}$ exactly when $x_k = 1$. Thus, for each $k$ with $x_k = 1$, we must have $v \cdot \sigma^{-k}x = 1$. Now, note that

$$v \cdot \sigma^{-k}x = \sum_{i:x_i=1} v_{i-k} = \sum_{i \in A} v_{i-k} = \sum_{i \in A} \mathbb{1}_{k \to i \text{ is an edge}}$$

which is exactly the outdegree of $k$ in $G_v[A]$ modulo 2. Thus, $G_v[A]$ is our required subgraph.

Conversely, suppose that $A$ is a set with odd size such that every vertex in $G_v[A]$ has odd outdegree. Define the vector $x$ by $x = \sum_{i \in A} e_i$. As $A$ has odd size, $\mathbf{e} \cdot \sigma^{-k}x = 0$ if and only if $k \in A$. But for such a $k$, $v \cdot \sigma^{-k}x = 1$ as it is the outdegree of $k$ in $G_v[A]$ modulo 2. Thus, no shift of $x$ is orthogonal to both $\mathbf{e}$ and $v$, proving that $\mathbf{e}$ and $v$ do not work together. □

*Remark.* Proposition 4.4.1 gives an easy proof of Proposition 4.3.5. Indeed, if $v$ is symmetric, then $G_v$ is a proper graph, and so, by the Handshaking Lemma, any subgraph with an odd number of vertices must have a vertex with even degree (in the subgraph).

By the following lemma, outdegree can also be replaced with indegree.

**Lemma 4.4.2.** *A Cayley digraph has an odd induced subgraph where each vertex has odd indegree if and only if it has an odd induced subgraph where each vertex has odd outdegree.*

*Proof.* If there is an edge from $i$ to $j$ in $G_v[A]$, then there is an edge from $-j$ to $-i$ in $G_v[-A]$. Hence, passing from $G_v[A]$ to $G_v[-A]$ swaps the indegrees and the outdegrees (and the number of vertices remains odd). □

We call an induced subgraph $H$ of $G_v$ a *bad subgraph* if

1. $H$ has an odd number of vertices and
2. at least one of the following holds:
   (a) each vertex in $H$ has odd outdegree within $H$;
   (b) each vertex in $H$ has odd indegree within $H$.

In particular, an odd induced cycle is a bad subgraph.

Proposition 4.4.1 and Lemma 4.4.2 imply that for $v \in \mathbb{F}_2^n$ with $v_0 = 0$, $v$ works with $\mathbf{e}$ if and only if $G_v$ has no bad subgraphs. An example of a bad subgraph is given in Figure 4.1.

In what follows, we will assume that certain non-symmetric vectors work together with $\mathbf{e}$ and will try to find a contradiction by finding a bad subgraph.

### 4.4.2 Restricting the edges of the digraph

For the remainder of the argument, we will work in $\mathbb{F}_2^p$, where $p$ is an odd prime. We will now begin searching for a bad subgraph in $G_v$, for $v \in \mathbb{F}_2^p$ a non-symmetric vector with $v_0 = 0$. Given such a $v$ that works with $\mathbf{e}$, $G_v$ is a directed graph which may include two-way edges. We will first show that we may assume that this does not happen, which means that $G_v$ is a proper digraph.

**Lemma 4.4.3.** *Suppose that $v \in \mathbb{F}_2^p$ is non-symmetric. Suppose that there are simultaneously $i$ with $v_i = v_{-i} = 1$ and $j \neq 0$ with $v_j = v_{-j} = 0$. Then $G_v$ has a bad subgraph.*

*Proof.* Suppose that $v \in \mathbb{F}_2^p$ works with $\mathbf{e}$; in other words, $G_v$ has no bad subgraphs. Define the sets $A = \{i : v_i = v_{-i} = 1\}$ and $B = \{j \neq 0 : v_j = v_{-j} = 0\}$. The condition on $v$ is therefore that both $A$ and $B$ are nonempty.

To prove Lemma 4.4.3, we will use the following claim.

**Claim.** $A + B \subseteq A \amalg B$.

*Proof of Claim.* Suppose that $i \in A$ and $j \in B$. We must show that $i + j \in A \amalg B$. Assume this is not the case and consider the subgraph of $G_v$ induced on $\{-i, 0, j\}$. By our assumptions, there is no edge between 0 and $j$, a two-way edge between 0 and $-i$ and a one-way edge between $-i$ and $j$.

There are two cases, depending on the direction of the edge between $-i$ and $j$. If it is directed from $-i$ to $j$, then each vertex has outdegree 1. Otherwise, each vertex has indegree 1. In either case, $G_v[\{-i, 0, j\}]$ is a bad subgraph. $\Diamond$

Now, we have $A + B \subseteq A \amalg B$, which means that $A_0 + B_0 \subseteq A \amalg B \amalg \{0\}$, where $A_0 = A \amalg \{0\}$ and $B_0 = B \amalg \{0\}$. We know that $|A_0| > 1$ and $|B_0| > 1$ since $A$ and $B$ are assumed to be nonempty. Furthermore, $|A_0 + B_0| \leqslant |A| + |B| + 1 \leqslant p - 2$, since otherwise $A \amalg B \amalg \{0\} = \mathbb{Z}/p\mathbb{Z}$ and $v$ is symmetric.

Thus, Vosper's Theorem allows us to conclude that $A_0$ and $B_0$ are both arithmetic progressions with the same common difference $d$. Since $A = -A$ and $B = -B$ are both symmetric, we deduce that $\{d, -d\}$ is contained in both $A_0$ and $B_0$, which contradicts the fact that they are disjoint. $\square$

If $v$ is a vector such that there is at least one value of $i$ with $v_i = v_{-i} = 1$, we will say that $v$ is *large*. Otherwise, $v$ is *small*. In particular, $v$ is small if and only if $G_v$ is a proper digraph. Lemma 4.4.3 tells us that if $v$ is non-symmetric and works with $\mathbf{e}$, either $v$ or $\mathbf{e} + v$ must be small.

Consequently, we will only consider small vectors $v$. Given such a $v$, we will start to look for bad subgraphs of $G_v$. We will do this by looking at the shortest directed cycle in $G_v$. This cycle must have length at least 3 because $G_v$ is a proper digraph, and must be induced else $G_v$ would have a shorter cycle. We will say that the *girth* of $G_v$, or (abusing notation) the *girth* of $v$, is the length of the shortest cycle in $G_v$.

For the results below, we assume the following set-up. Let $v \in \mathbb{F}_2^p$ be a nonzero small vector with $v_0 = 0$ that works together with $\mathbf{e}$. Let $A = A^{(v)}$ be the set $\{i : v_i = 1\}$. Since $v$ is small, $A$ and $-A$ are disjoint. Let $A_0^{(v)} = A_0$ be $A \amalg \{0\}$ and let $k$ be the girth of $G_v$. We may immediately make some simple observations.

**Lemma 4.4.4.** *With $v, G_v, A, A_0$ and $k$ as defined above, the following hold.*

1. *$k$ is even.*

2. *$0 \notin (k-2)A_0 + A$, or in other words*

$$|(k-2)A_0 + A| \leqslant p - 1. \tag{4.4.1}$$

*Proof.*

1. The shortest cycle in $G_v$ is an induced subgraph in which each vertex has degree 1. Thus, as $v$ and $\mathsf{e}$ work together, such a subgraph must have an even number of vertices, which means that $k$ must be even.

2. If 0 can be written as a sum of $k-1$ elements of $A_0$, not all zero, then we get a corresponding cycle of length at most $k-1$ in $G_v$, which we assume does not exist. The inequality (4.4.1) follows immediately. □

The next proposition gives bounds on the size of $A$.

**Proposition 4.4.5.** *With $v, G_v, A, A_0$ and $k$ as defined above,*

$$\frac{p}{k} < |A| < \frac{p}{k-1}. \tag{4.4.2}$$

The proof of Proposition 4.4.5 relies on the following lemma.

**Lemma 4.4.6.** *Suppose that $C$ is a $k$-cycle in $G_v$. For any $i \notin C$, there are at least two edges between $i$ and $C$ (ignoring direction).*

*Proof.* If there is one edge from $i$ to $C$, then $C \amalg \{i\}$ induces a bad subgraph as each vertex has outdegree 1. If there is one edge from $C$ to $i$, then $C \amalg \{i\}$ induces a bad subgraph as each vertex has indegree 1. In either case, we find a bad subgraph, contradicting the assertion that $v$ and $\mathsf{e}$ work together.

Suppose instead that there are no edges at all between $i$ and $C$. Without loss of generality (since $G_v$ is vertex transitive), we may assume $0 \in C$ and write $C$ as $C = \{0, a_1, a_1+a_2, \ldots, a_1+\cdots+a_{k-1}\}$ for some $a_1, \ldots, a_k \in A$ with $a_1+\cdots+a_k = 0$.

Consider doing a single *swap* of $a_1$ with $a_2$; that is, we replace the order of the edges labelled $a_1$ and $a_2$ in the cycle $C$ to get a new cycle $C_1$. Thus, $C_1 = C \setminus \{a_1\} \cup \{a_2\}$. Observe that the subgraph induced on $C_1$ is also a $k$-cycle, as any other edges between vertices would create a shorter cycle, and this does not exist by the assumption that $C$ is a shortest cycle. Note that this swap changes the number of edges between $i$ and the cycle by at most one, since at most one vertex is changed. There cannot be an edge between the new vertex $a_2$ and $i$ since that would put us in the earlier case, so we find that $i$ is not connected to $C_1$.

We can continue doing this; swap $a_1$ with $a_3$ to form $C_2 = \{0, a_2, a_2+a_3, a_1+a_2+a_3, \ldots, a_1+\cdots+a_{k-1}\}$, and again $i$ can have no edges to or from $C_2$. After $k-1$ iterations, we form the cycle $C_{k-1} = \{0, a_2, a_2+a_3, \ldots, a_2+a_3+\cdots+a_k\}$, and there are no edges between $i$ and $C_{k-1}$. Since $\sum_i a_i = 0$, we can rewrite this as

$C_{k-1} = \{a_1 - a_1, a_1 + a_2 - a_1, \ldots, a_1 + a_2 + \cdots + a_{k-1} - a_1, 0 - a_1\}$ so $C_{k-1} = C - a_1$ (having shifted any fixed starting point).

If we consider $C_{k-1}$ as starting at $-a_1$, the first edge to 0 again corresponds to adding $a_1$ and we can repeat this procedure, but starting with $C_{k-1}$. This gives the cycle $C_{2(k-1)} = C - 2a_1$ and we can continue iterating. Since $p$ is prime and $a_1 \neq 0$, we will see the family of cycles $\{C - j : j \in \mathbb{Z}/p\mathbb{Z}\}$, and, in particular, every point is in some cycle.

However, throughout this procedure, $i$ will never be connected to any point in any of these cycles. This contradicts the assertion that every element of $\mathbb{Z}/p\mathbb{Z}$ is contained in some cycle, in view of the fact that $i$ is connected to every element of $i + A$.

Thus, there must be at least two edges between $i$ and $C$. □

*Proof of Proposition 4.4.5.* The upper bound follows immediately from (4.4.1) and $k - 2$ applications of the Cauchy-Davenport inequality.

For the lower bound, we will use Lemma 4.4.6. Let $C$ be the vertex set of a cycle of length $k$ in $G_v$. We will count the number of pairs $(i, j)$ with $i \in \mathbb{Z}/p\mathbb{Z}$ and $j \in C$, such that there is an edge between $i$ and $j$. Each element of $C$ has $|A|$ outedges and $|A|$ inedges. Thus, there are $2|A|$ pairs for each $j \in C$, and so $2k|A|$ pairs in total.

On the other hand, Lemma 4.4.6 tells us that there are at least 2 edges for each $i \in \mathbb{Z}/p\mathbb{Z}$, and so there are at least $2p$ edges. Therefore, $2p \leqslant 2k|A|$, which rearranges to give the lower bound claimed in Proposition 4.4.5.

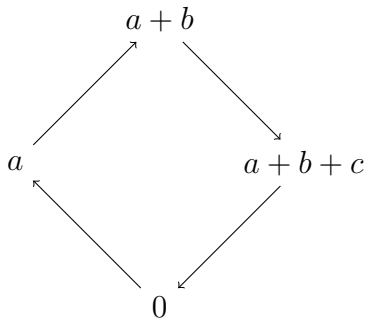
To see the strict inequality, observe that $k$ is not a factor of $p$; indeed, if $k = p$ then $C = \mathbb{Z}/p\mathbb{Z}$ induces an odd directed cycle. □

### 4.4.3 Eliminating large girth

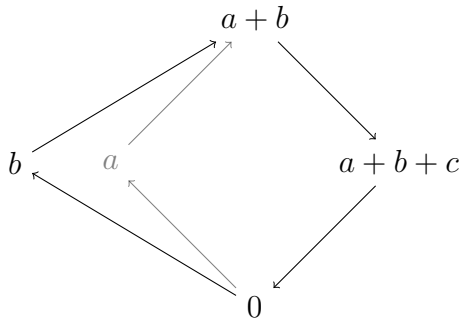
We will continue to restrict the possible small non-symmetric vectors $v$ that can work together with $\mathbf{e}$. Recall that $A = \{i : v_i = 1\}$ and $A_0 = A \amalg \{0\}$, and since $v$ is small, $A$ and $-A$ are disjoint. Denote again $G_v$ for the Cayley graph generated by $A$ as defined before Proposition 4.4.1. The main result of this subsection is an upper bound on the girth $k$ of $G_v$.

**Proposition 4.4.7.** *If a small $v \neq 0$ works together with $\mathbf{e}$, then the girth of $v$ is 4 or 6.*

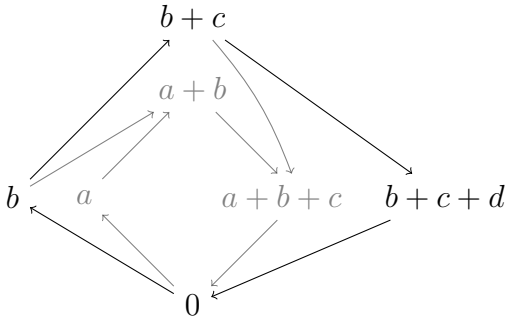
Our proof of Proposition 4.4.7 will rely on a weak version of a theorem due to Grynkiewicz.



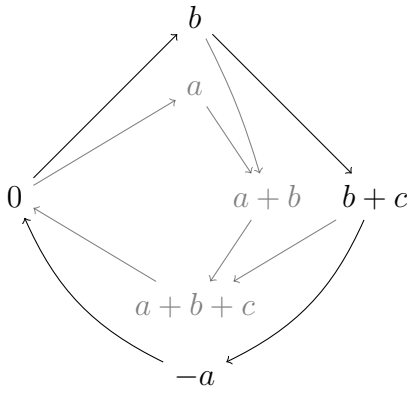

(a) The induced 4-cycle $C$.




(b) The cycle $C_1$ formed by swapping $a$ with $b$.




(c) The cycle $C_3 = C - a$.




(d) The cycle $C_3 = C - a$ after changing the point of view.

Figure 4.2: Various snapshots of $C$ throughout the swapping process.

**Theorem 4.4.8** ([30, Theorem 19.3])**.** *Let $p$ be a prime, let $\alpha \in (0, 0.45695]$ and let $X \subseteq \mathbb{Z}/p\mathbb{Z}$ be a nonempty subset such that $|2X| = 2|X| - 1 + r < p$. Suppose that the following two bounds hold:*

$$|2X| \leqslant (2+\alpha)|X| - 3, \tag{4.4.3}$$

$$|2X| \leqslant \min\left\{\frac{p+3}{2}, \frac{9-(1+2\alpha)^2(2+\alpha)}{9-(1+2\alpha)^2}p\right\}. \tag{4.4.4}$$

*Then, $X$ is contained in an arithmetic progression of length at most $|A| + r$.*

*Remark.* Theorem 4.4.8 is similar to Freiman's $3k-3$ theorem [54, Theorem 5.11], except in $\mathbb{Z}/p\mathbb{Z}$ rather than $\mathbb{Z}$ (and with slightly altered conditions). In Freiman's $3k-3$ theorem, the conditions (4.4.3) and (4.4.4) are not needed; instead, the conclusion is satisfied by any set $A$ with $|2A| < 3|A| - 3$. It is conjectured that a similar weakening of the conditions should hold in $\mathbb{Z}/p\mathbb{Z}$ (see [30, Conjecture 19.2]).

There are various similar results in the literature which also provide weak forms of Freiman's $3k-3$ theorem in $\mathbb{Z}/p\mathbb{Z}$; for instance, Serra and Zémor prove in [51] that the size constraint (4.4.4) can be omitted, at the expense of a worse bound on the doubling $|2X| \leqslant (2+\varepsilon)|X|$ for some small $X$ (and sufficiently large $p$). Such a result could be used to obtain a version of Proposition 4.4.7 with a weaker bound on the girth.

In particular, we obtain the following instance of Theorem 4.4.8 by taking $\alpha = 0.36$.

**Corollary 4.4.9.** *Let $p$ be a prime, and let $X \subseteq \mathbb{Z}/p\mathbb{Z}$ be a nonempty subset. Suppose that*

$$|2X| \leqslant 2.36|X| - 3 \tag{4.4.5}$$

$$|2X| \leqslant 0.334p. \tag{4.4.6}$$

*Then $X$ is contained in an arithmetic progression with $|2X| - |X| + 1$ terms.*

We will also need a strong upper bound on the value of $k$, depending on $p$, in the case that $A_0$ is not an arithmetic progression.

**Lemma 4.4.10.** *If $A_0$ is not an arithmetic progression, then $k \leqslant \sqrt{p+2}$.*

*Proof.* Suppose that $A_0$ is not an arithmetic progression. Note that this implies $|A| \geqslant 2$.

Equation (4.4.1) tells us that $|(k-2)A_0 + A| \leqslant p-1$. The Cauchy-Davenport inequality gives $|(k-2)A_0| \leqslant p-|A|$. Applying the contrapositive of Vosper's Theorem (see Theorem 4.3.8) to $|(k-3)A_0+A_0| \leqslant p-2$ gives that $|(k-2)A_0| > |(k-3)A_0|+|A_0|$. Applying this $k-4$ further times, we conclude that $(k-2)|A_0| \leqslant |(k-2)A_0| \leqslant p-|A|$.

Write $p = rk+s$, where $0 < s < k$. We have that $|A| \geqslant r+1$ by Proposition 4.4.5, and so $(k-2)(r+2) \leqslant (rk+s)-(r+1)$, which rearranges to give

$$2k-3 \leqslant r+s. \tag{4.4.7}$$

If $r < k-1$, then this equation gives $r = k-2$, $s = k-1$, and hence $p = k^2-k-1$. The only value of $|A|$ satisfying Proposition 4.4.5 is $|A| = k-1$, which is odd because $k$ is even. However, $|A|$ must be even, else the whole of $G_v$ is a bad subgraph, so we must have $r \geqslant k-1$. If $r \geqslant k$, we have $p \geqslant k^2$. Else we have $r = k-1$ and equation 4.4.7 gives $s \geqslant k-2$, and so $p \geqslant k^2-2$. □

Finally, we will need to use the following theorem due to Hamidoune and Rødseth [31], which is an extension of Vosper's Theorem. Say that an *almost progression* is an arithmetic progression, possibly missing one term.

**Theorem 4.4.11** ([31, Theorem 3])**.** *Suppose that $A, B \subseteq \mathbb{Z}/p\mathbb{Z}$ are sets which satisfy $|A|, |B| \geqslant 3$. Suppose further that*

$$7 \leqslant |A+B| \leqslant |A|+|B| \leqslant p-4.$$

*Then $A$ and $B$ are both almost progressions with the same common difference.*

We are now ready to prove Proposition 4.4.7.

*Proof of Proposition 4.4.7.* Suppose that $v$ is a vector with girth $k$ that works with $\mathbf{e}$. Suppose further that $k$ is not 4 or 6; this implies $k \geqslant 8$ since $k$ is even.

We will start by using Corollary 4.4.9 and Theorem 4.4.11 to prove the following claim.

**Claim.** *$A_0$ is contained in an arithmetic progression of length at most $\frac{p}{k-2}+1$.*

*Proof of Claim.* Our proof will fall into three cases, depending on the structure of $A_0$ and $2A_0$.

**Case 1:** *$A_0$ is already an arithmetic progression.* In this case, we may use the upper bound of Proposition 4.4.5 to see that

$$|A_0| < \frac{p}{k-1}+1 < \frac{p}{k-2}+1.$$

Thus, we may take the progression $P$ to be $A_0$, and it will already be of the required length.

**Case 2:** *$A_0$ is not an arithmetic progression, but $2A_0$ is an almost progression.* First, we will use (4.4.1) to obtain an upper bound on $|2A_0|$. Indeed, one application of Cauchy-Davenport tells us that $|(k-2)A_0| \leqslant p - |A|$, a second application gives that

$$\left|\left(\frac{k-4}{2}\right)(2A_0)\right| + |2A_0| - 1 \leqslant |(k-2)A_0|,$$

and a further $(k-6)/2$ applications give

$$\frac{k-2}{2}(|2A_0| - 1) \leqslant |(k-2)A_0|.$$

Since $|A| > p/k$ by Proposition 4.4.5, this immediately tells us that

$$|2A_0| < \frac{2p(k-1)}{k(k-2)} + 1. \tag{4.4.8}$$

Now, suppose without loss of generality that $2A_0$ is an almost progression with common difference 1, which we will treat as a subset of $\mathbb{Z}$ in the natural way. Say that the endpoints of $2A_0$ are $-r$ and $s$, for $r, s \geqslant 0$. In particular, $-r$ and $s$ are both elements of $2A_0$.

We immediately obtain that

$$r + s \leqslant \frac{2p(k-1)}{k(k-2)} + 1, \tag{4.4.9}$$

since there is at most one element of $[-r, s]$ missing from $2A_0$.

Now, we claim that $A_0 \subseteq [-\lfloor r/2 \rfloor, \lfloor s/2 \rfloor]$. Indeed, $0 \in A_0$, so for any $x \in A_0$, $x + 0 \in 2A_0 \subseteq [-r, s]$. Therefore, $A_0 \subseteq [-r, s]$.

However, if $x \in A_0$, then $2x \in [-r, s]$. If $x \in [-r, -\lfloor r/2 \rfloor - 1]$, then $2x \in [-2r, -r-1]$. Because $p - 2r > s$, which follows from (4.4.9) and the fact that $k \geqslant 8$, it follows that $A_0$ does not meet $[-r, -\lfloor r/2 \rfloor - 1]$. A similar argument rules out $[\lfloor s/2 \rfloor + 1, s]$, and so it must be the case that $A_0 \subseteq [-\lfloor r/2 \rfloor, \lfloor s/2 \rfloor]$ as claimed.

Thus, $A_0$ is contained in a progression of length at most $\lfloor r/2 \rfloor + \lfloor s/2 \rfloor + 1$, which is at most

$$\frac{r+s}{2} + 1 \leqslant \frac{p(k-1)}{k(k-2)} + \frac{3}{2}.$$

This will be no greater than $p/(k-2) + 1$ provided that $p \geqslant k(k-2)/2$, which follows from Lemma 4.4.10.

**Case 3:** $2A_0$ *is not an almost progression.* Our plan is to prove that $A_0$ obeys the conditions (4.4.5) and (4.4.6). Again, we may use (4.4.1) to show that $|(k-2)A_0| \leqslant p - |A| < p(k-1)/k$.

Observe that $|2A_0| \geqslant 3$ and $|4A_0| \geqslant 7$, both of which follow from the assertion that $|A_0| \geqslant 3$. Furthermore, $|(k-2)A_0| \leqslant p(k-1)/k$, which is less than $p-4$ by Lemma 4.4.10 and the fact that $k \geqslant 8$. Thus, we may apply the contrapositive of Theorem 4.4.11 $(k-4)/2$ times to show that

$$|(k-2)A_0| \geqslant \frac{k-2}{2}|2A_0| + \frac{k-4}{2},$$

which, combined with our upper bound $|(k-2)A_0| \leqslant p - |A| < p(k-1)/k$, gives

$$|2A_0| \leqslant \frac{2p(k-1)}{k(k-2)} - \frac{k-4}{k-2}. \tag{4.4.10}$$

Now, we will first prove that $A_0$ satisfies (4.4.6). Indeed, this is the case provided that

$$\frac{2p(k-1)}{k(k-2)} - \frac{k-4}{k-2} \leqslant 0.334p.$$

Since $k \geqslant 8$, the left hand side is at most $7p/24$, which is less than $0.334p$.

It remains to prove that $A_0$ satisfies (4.4.5). Suppose that this is not the case; in other words,

$$|2A_0| > 2.36|A_0| - 3. \tag{4.4.11}$$

In view of (4.4.10) and Proposition 4.4.5, we deduce that

$$\frac{2p(k-1)}{k(k-2)} - \frac{k-4}{k-2} > 2.36\left(\frac{p}{k}+1\right) - 3.$$

Rearranging, this becomes

$$p\left(\frac{0.36k - 2.72}{k(k-2)}\right) < 0.64 - \frac{k-4}{k-2}.$$

However, when $k \geqslant 8$ the left hand side of this is positive and the right hand side is negative, which gives a contradiction.

Thus, $A_0$ satisfies (4.4.5) and (4.4.6), and so we may apply Corollary 4.4.9. This tells us that $A_0$ is contained in an arithmetic progression with $|2A_0| - |A_0| + 1$ terms, and this is at most $p/(k-2) + 1$ by Proposition 4.4.5 and the bound (4.4.10). $\Diamond$

Without loss of generality, we may assume $A_0$ is contained in an arithmetic progression $P$ of common difference 1, and that both endpoints of $P$ are elements of $A_0$.

We will derive a contradiction by finding a bad subgraph. We will split into two cases depending on whether or not 0 is an endpoint of $P$.

**Case A:** 0 *is not an endpoint of* $P$. We will treat the elements of $P$ as elements of $\mathbb{Z}$ in the natural way.

Let $P = [-R, S]$, where $R \neq S$ follows from the fact that $A$ and $-A$ are disjoint. Without loss of generality, assume that that $R < S$. The bound $|P| \leqslant \frac{p}{k-2} + 1$ tells us that $R + S \leqslant \frac{p}{k-2}$. Let $-r$ be the largest negative element of $A_0$ and $s$ be the smallest positive element of $A_0$. Consider $H = G_v[[-r, s]]$, the subgraph of $G_v$ induced on the interval $[-r, s]$.

Each nonnegative vertex of $H$ has an outgoing edge corresponding to $-r$, and each nonpositive vertex of $H$ has an outgoing edge corresponding to $+s$. Consequently, each vertex of $H$ has positive outdegree. In particular, $H$ contains a cycle $C$, which we may assume to be induced by taking the shortest such cycle. In other words, $G_v$ contains an induced cycle $C$ contained in $[-r, s]$.

If $C$ is an odd cycle, then it is the required bad subgraph. Otherwise, $C$ must be even. Translate so that $-r$ is the smallest element of $C$. We claim that $-r - S$ has exactly one outgoing edge pointing towards $C$, and no incoming edges from $C$. Indeed, there is an edge directed from $-r - S$ to $-r$ because $S$ is assumed to be in $A$.

Now, the outgoing edges from $-r - S$ go to $-r - S + A$, which is contained in $-r - S + [-R, S]$, and the incoming edges to $[-r - S]$ come from $-r - S - A$, which is contained in $-r - S - [-S, R]$. Because $S > R$, we have that

$$(-r - S + A) \cup (-r - S - A) \subseteq -r - S + [-S, S],$$

and so there will be no other edges between $-r - S$ and $C$ provided that $(-r - S + [-S, S]) \cap [-r, s] = \{-r\}$.

To see that this is the case, observe that $-r - S + [-S, S] = [-r - 2S, -r]$. Working in $\mathbb{Z}/p\mathbb{Z}$, the intervals $[-r - 2S, -r - 1]$ and $[-r, s]$ will be disjoint provided that $p - r - 2S > s$. This holds because $r + s$ and $S$ are both less that $\frac{p}{k-2} < \frac{p}{3}$.

Thus, there is exactly one outgoing edge from $-r - S$ to $C$, and no incoming edges from $C$ to $-r - S$. This means that each vertex in $C \amalg \{-r - S\}$ has outdegree 1, and so this is our required bad subgraph.

**Case B:** 0 *is an endpoint of* $P$. We will treat elements of $P$ as elements of $[0, p]$ in the natural way. Let $m = \max P$, so $m < \frac{p}{k-2}$.

We claim that any $r$ with the property that $m/2 \leqslant r \leqslant 3m/2$ is the sum of two elements of $A_0$. Indeed, $|A_0| > p/k + 1$, so there are fewer than $m - p/k - 1$ elements

of $P \setminus A_0$. However, given any $r$ with $m/2 \leqslant r \leqslant 3m/2$, there are at least $m/4$ disjoint pairs of elements of $P$ which sum to $r$. Thus, if $r$ is not expressible as the sum of two elements of $A$, then it must be the case that

$$\frac{m}{4} < m - \frac{p}{k} - 1.$$

Since $m < p/(k-2)$ and $k \geqslant 8$, this cannot happen.

Now, choose $r$ with $m/2 \leqslant r \leqslant 3m/2$ so that $p - r$ is divisible by $m$ (this is possible since there are at least $m$ integers in the interval $[m/2, 3m/2]$), and let $r_1$ and $r_2$ be two elements of $A_0$ which sum to $r$. This gives two cases depending on whether $r_1$ and $r_2$ are both nonzero or at least one is zero; we prove only the first case, as the second is easier.

We write $p = sm + r_1 + r_2$; since $v$ has girth at least 8, we must have $s \geqslant 6$. If $s$ is odd, then we have an odd induced cycle of length $s+2$ whose elements are the partial sums of

$$\underbrace{m + \cdots + m}_{s-1 \text{ terms}} + r_1 + m + r_2 = p.$$

Note that the order of the elements is important, and that $m$ is chosen (and placed in between $r_1$ and $r_2$) such that the odd cycle

$$\{0, m, 2m, \ldots, (s-1)m, (s-1)m + r_1, sm + r_1\}$$

is induced.

Otherwise, $s$ is even. By a similar argument to the one used to show that $r$ is a sum of two elements of $A_0$, we can deduce that $m$ is the sum of two elements of $A$; we write $m = m_1 + m_2$. This gives us an odd induced cycle of length $s+3$, whose elements are the partial sums of

$$m + m_1 + m + m_2 + \underbrace{m + \cdots + m}_{s-4 \text{ terms}} + r_1 + m + r_2 = p. \qquad \square$$

## 4.5 Proof of the main result

Unfortunately, we were unable to show that there are no small vectors $v \in \mathbb{F}_2^p$ with girth at most 6 which work together with $\mathbf{e}$. However, in order to prove that $h_2(p) = 2$, it suffices to show that there are no two vectors $v$ and $w$, such that $\mathbf{e}$, $v$ and $w$ are linearly independent and work together.

Suppose that such a pair of vectors exists. As before, set $A^{(v)} = \{i : v_i = 1\}$ and $A_0^{(v)} = A^{(v)} \amalg \{0\}$, and define $A^{(w)}$ and $A_0^{(w)}$ similarly. Recall from Lemma 4.4.3 that

we may assume that $v$ and $w$ are small, possibly by replacing with $\mathbf{e}+v$ or $\mathbf{e}+w$ if necessary. In other words, $A^{(v)}$ is disjoint from $-A^{(v)}$, and the same holds for $A^{(w)}$.

Since $v, w$ and $\mathbf{e}$ all work together, $v+w$ must also work with $\mathbf{e}$. Now, $v+w$ cannot be symmetric, else the fact that $\mathbf{e}, v$ and $v+w$ work together would contradict Lemma 4.3.6. Thus, $v+w$ must be small or large. We will now show that the latter of these cannot occur.

**Lemma 4.5.1.** *Suppose that $v, w \in \mathbb{F}_2^p$ are small vectors such that $v, w$ and $\mathbf{e}$ work together. Then, the vector $v+w$ must also be small.*

*Proof.* Suppose otherwise that $v+w$ is large. Since $v$ and $w$ are small, they must have girth at least 4, and so, by Proposition 4.4.5, $A^{(v)}$ and $A^{(w)}$ have size at most $\lfloor p/3 \rfloor$. Since $A^{(v+w)} = A^{(v)} \triangle A^{(w)}$, $A^{(v+w)}$ must have size at most $2\lfloor p/3 \rfloor$, which means that $|A^{(\mathbf{e}+v+w)}| \geqslant p - 1 - 2\lfloor p/3 \rfloor$.

However, by assumption $\mathbf{e}+v+w$ is a small vector that works with $\mathbf{e}$, and so has girth at least 4. In particular, $|A^{(\mathbf{e}+v+w)}| \leqslant \lfloor p/3 \rfloor$.

Combining these two equations, we find that $(p-1)/3 \leqslant \lfloor p/3 \rfloor$, which implies that, for some $m \geqslant 2$, $p = 3m+1$ and $|A^{(v)}| = |A^{(w)}| = |A^{(\mathbf{e}+v+w)}| = m$.

To rule out this special case, we will use Vosper's Theorem to find the structure of $v$ and show that $G_v$ has an induced 5-cycle, which contradicts the assertion that $v$ works with $\mathbf{e}$. By two applications of Cauchy-Davenport, $|2A_0^{(v)} + A^{(v)}| \geqslant 3m$. However, $|2A_0^{(v)} + A^{(v)}| \leqslant 3m$ by (4.4.1). Therefore, equality holds, and we may apply Vosper's Theorem to see that $A_0^{(v)}$ and $A^{(v)}$ are arithmetic progressions with the same common difference; without loss of generality, the common difference is 1, and $A_0^{(v)} = \{0, \ldots, m\}$.

Now, we have an induced 5-cycle whose vertices are $\{0, m-1, m+1, 2m, 2m+2\}$. Indeed, $m-1$ and 2 are both elements of $[m]$, and there will be no other edges provided that $2m-2 > m$, which holds for $p > 7$. □

As with Proposition 4.4.1, we can rephrase our problem in terms of a graph theoretical problem. Since $v+w$ is small, there cannot be any edges that have different directions in $G_v$ and $G_w$. Thus, we may encode $G_v$ and $G_w$ in a single graph $G$ with vertex set $\mathbb{Z}/p\mathbb{Z}$ using edges coloured with three colours. We do this as follows.

- If there is an edge from $i$ to $i+j$ in both $G_v$ and $G_w$ (so $j \in A^{(v)} \cap A^{(w)}$), then draw a green edge from $i$ to $i+j$.
- If there is an edge from $i$ to $i+j$ in $G_v$ but not in $G_w$ (so $j \in A^{(v)} \setminus A^{(w)}$), then draw a red edge from $i$ to $i+j$.

- Finally, if there is an edge from $i$ to $i+j$ in $G_w$ but not in $G_v$ (so $j \in A^{(w)} \setminus A^{(v)}$), then draw a blue edge from $i$ to $i+j$.

Thus, $G_v$ consists of the red and green edges of $G$, and $G_w$ consists of the blue and green edges. Furthermore, $G_{v+w}$ consists of the red and blue edges.

The relevance of $G$ to our problem is given by the following proposition, which is analogous to Proposition 4.4.1.

**Proposition 4.5.2.** *If $v, w \in \mathbb{F}_2^p$ are two small vectors with $v_0 = w_0 = 0$, then $v, w$ and $\mathbf{e}$ work together if and only if there is no induced subgraph $H$ of $G$ on an odd number of vertices such that, for each $i \in H$, the numbers of red, blue and green outneighbours of $i$ in $H$ are not all of the same parity.*

As with Lemma 4.4.2, the outdegrees can be replaced with indegrees. Thus, for the remainder of the proof, we will call an induced subgraph $H$ of $G$ *bad* if

1. $H$ has an odd number of vertices and
2. at least one of the following holds:
   (a) for each vertex in $H$, the outdegrees in the three colours do not all have the same parity;
   (b) for each vertex in $H$, the indegrees in the three colours do not all have the same parity.

In other words, Proposition 4.5.2 implies that $v$ and $w$ work together with $\mathbf{e}$ if and only if $G$ has no bad subgraphs.

*Proof of Proposition 4.5.2.* The vectors $v$, $w$ and $\mathbf{e}$ fail to work together if and only if there is a vector $x$ for which any shift $\sigma^k x$ is not orthogonal to at least one of them. Let $B = \{i : x_i = 1\}$.

As shown in the proof of Proposition 4.4.1, we find that $v \cdot \sigma^{-k} x = 1$ if and only if $k$ has odd outdegree in $G_v[B]$, and similarly with $v$ replaced by $w$.

Thus, $v, w$ and $\mathbf{e}$ fail to work together if and only if there is a subset $B$ such that for each $k \in B$, $k$ has odd outdegree in at least one of $G_v[B]$ and $G_w[B]$. Since $G_v$ consists of the red and green edges of $G$, the outdegree of some $k \in B$ is even in $G_v[B]$ if and only if the number of red and green outedges of $k$ in $G[B]$ has the same parity. The same statement holds of for $w$ with red replaced by blue. Hence the statement that $k$ has odd outdegree in at least one of $G_v[B]$ and $G_w[B]$ is equivalent to the assertion that the outdegrees of $k$ in $G[B]$ in the three colours are not all of the same parity. □

**Observation 4.5.3.** *For any colouring of the edges, a directed triangle is a bad subgraph.*

Let us write $A$ for $A^{(v)} \cup A^{(w)}$ so there is an edge from $i$ to $i+j$ in $G$ if and only if $j \in A$. Let $A_{\text{green}} = A^{(v)} \cap A^{(w)}$ (so there is a green edge from $i$ to $i+j$ if and only if $j \in A_{\text{green}}$) and similarly define $A_{\text{red}} = A^{(v)} \setminus A^{(w)}$ and $A_{\text{blue}} = A^{(w)} \setminus A^{(v)}$.

Since $v, w$ and $v+w$ all work with $\mathbf{e}$, Proposition 4.4.7 tells us that they must each have girth 4 or 6. To complete the proof of Theorem 4.1.1, we will divide into cases depending on how many of $v, w$ and $v+w$ have girth 4 and how many have girth 6.

**Lemma 4.5.4.** *Suppose that at least two of $v, w$ and $v+w$ have girth 4. Then $G$ has a bad subgraph.*

*Proof.* Without loss of generality, we can assume that both $v$ and $w$ have girth 4. By Proposition 4.4.5, $|A^{(v)}|, |A^{(w)}| > p/4$ and $|A^{(v+w)}| > p/6$. Since $A^{(v+w)} = A^{(v)} \triangle A^{(w)}$, each element $j \in A$ is in exactly 2 of $A^{(v)}$, $A^{(w)}$ and $A^{(v+w)}$. Hence,

$$\begin{aligned} |A| &= \frac{1}{2}\left(|A^{(v)}| + |A^{(w)}| + |A^{(v+w)}|\right) \\ &> \frac{1}{2}\left(\frac{p}{4} + \frac{p}{4} + \frac{p}{6}\right) = \frac{p}{3}. \end{aligned}$$

Thus, Cauchy-Davenport tells us that $|2(A \amalg \{0\}) + A| \geqslant p$, and in particular 0 is the sum of at most 3 elements of $A$. Thus, there is a cycle of length at most 3 in $G$. There are no two-way edges in $G$, so $G$ contains a directed triangle. □

For the remaining two cases, we will require the following lemma.

**Lemma 4.5.5.** *If $G$ contains a 4-cycle which includes edges of all three colours, then $G$ contains a bad subgraph.*

*Proof.* Suppose that $G$ contains a 4-cycle which includes edges of all three colours. Label a red edge $r$ (so the edge is directed from some $i$ to $i+r$), label a blue edge $b$, a green edge $g$ and the final edge $x$. Without loss of generality we can assume that the edges are in the order $r, g, b, x$. We will consider subgraphs induced on (subsets of) $\{0, r, g, b, r+g, r+b, b+g, r+g+b\}$.

First, consider the subgraph $H = G[\{0, r, g, r+g, r+g+b\}]$, as illustrated by Figure 4.4. Observe that there can be no edge between 0 and $r+g$. If there were an edge from 0 to $r+g$, then $\{0, r+g, r+g+b\}$ would induce a directed triangle, and if there were an edge from $r+g$ to 0 then $\{0, g, r+g\}$ would induce a directed

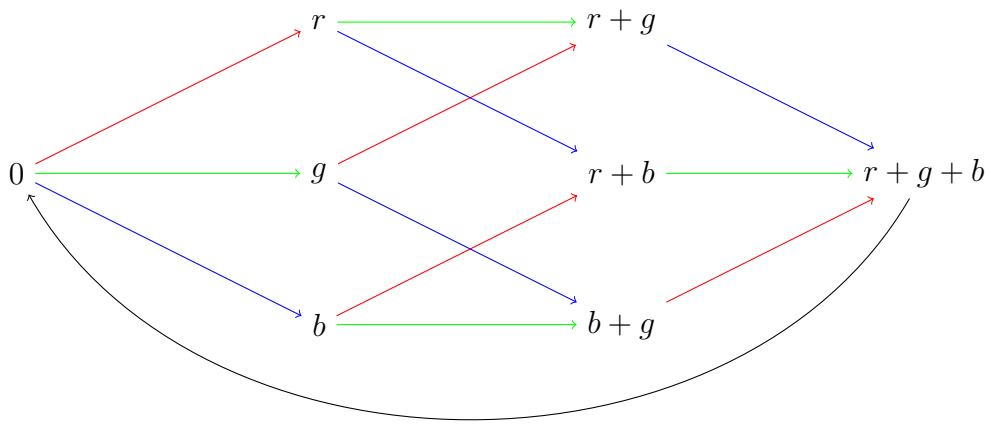



Figure 4.3: A rainbow 4-cycle implies the existence of a subgraph as depicted in the figure. The black edge can be coloured either red, green or blue.

triangle. In either case, we obtain a bad subgraph. Similarly, there can be no edge between $r+g+b$ and either $r$ or $g$. Thus, the only possible edge that can be added is between $r$ and $g$.

Now, observe that if there is no edge between $r$ and $g$, then $H$ is a bad subgraph. In fact, the only way to stop $H$ from being bad is to have a red edge from $g$ to $r$, or a green edge from $r$ to $g$.

The same argument can be applied to subgraphs induced by $\{0, r, b, r+b, r+g+b\}$ and $\{0, g, b, b+g, r+g+b\}$ to show that there is either a blue edge from $r$ to $b$ or a red edge from $b$ to $r$, and either a blue edge from $g$ to $b$ or a green edge from $b$ to $g$. If we take edges of all three colours, then $\{r, g, b\}$ would induce a directed triangle. Thus, one colour appears twice between pairs of $r, g$ and $b$; without loss of generality, that colour is red. We may hence assume there are red edges directed from $g$ and $b$ to $r$.

Now, consider the subgraph $H$ induced by the vertices $\{0, r, b, r+g, r+g+b\}$, as seen in Figure 4.5.

As before, the only edge that is undetermined is a possible edge between $b$ and $r+g$; such an edge cannot be directed from $r+g$ to $b$, else $\{b, r, r+g\}$ induces a triangle.

If there is no edge from $b$ to $r+g$, then $H$ is bad. Indeed, $H$ is bad unless there is a red edge from $b$ to $r+g$.

Finally, consider the subgraph induced on $\{0, b, g, r+g, r+b+g\}$ as depicted in Figure 4.6.

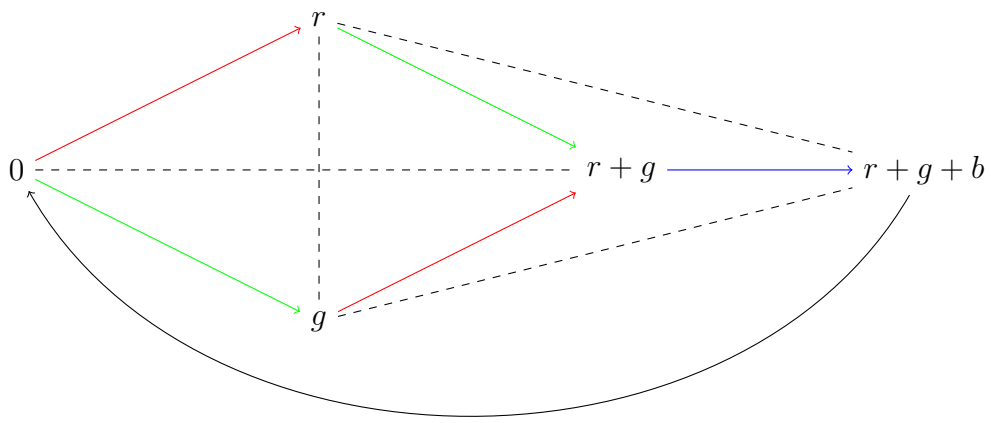



Figure 4.4: If none of the dotted edges is present, then the depicted subgraph of $G$ is a bad subgraph.

We know that there is an edge between $b$ and $g$ and that edge is either blue or green. In both cases, this is the bad subgraph we sought. □

**Lemma 4.5.6.** *Suppose that $v, w$ and $v + w$ all have girth 6. Then $G$ has a bad subgraph.*

*Proof.* As in the proof of Lemma 4.5.4, we learn that $|A| > p/4$, and so, by 3 applications of Cauchy-Davenport, $G$ contains a cycle of length at most 4. Since there can be no directed triangles (Observation 4.5.3), the girth must be at least 4, so there is an induced 4-cycle $C$.

Suppose that $C$ contains only red and green edges; then $C$ is contained entirely within $G_v$, contradicting the fact that $v$ has girth 6. Therefore, $C$ must contain at least one blue edge. Similarly $C$ must contain at least one red edge and at least one green edge. Thus, $C$ must contain all three colours, and so we are done by Lemma 4.5.5. □

**Lemma 4.5.7.** *Suppose that $v$ and $w$ have girth 6, and $v + w$ has girth 4. Then $G$ has a bad subgraph.*

*Proof.* As in the proof of Lemma 4.5.6, since $G_v$ and $G_w$ do not contain a 4-cycle, we find that there cannot be a 4-cycle in $G$ with no blue edges or with no red edges. Since there cannot be a 4-cycle with all three colours by Lemma 4.5.5, we find that $G$ cannot have a 4-cycle containing a green edge. Note that $G$ can also not contain an induced directed 3-cycle or 5-cycle, as such a subgraph is bad.

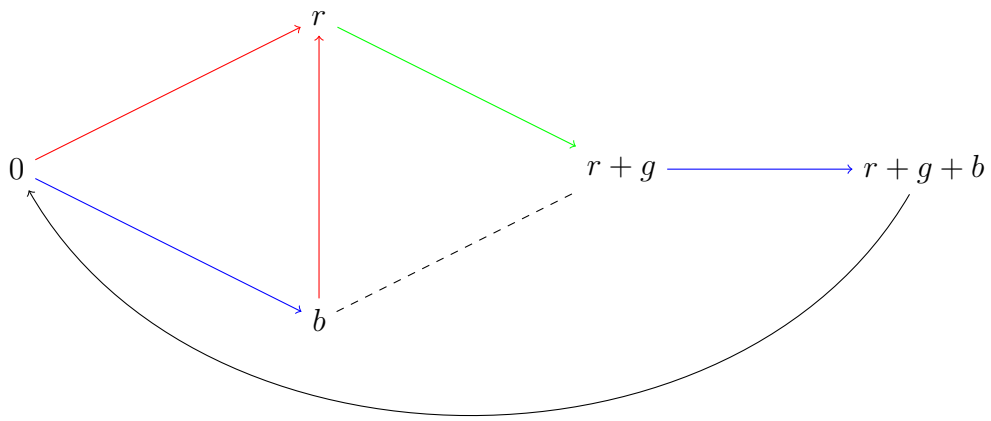



Figure 4.5: Vertex $r$ has an even number of edges coming in for each colour. If the dotted line is not an edge, then the depicted subgraph is bad on outdegrees.

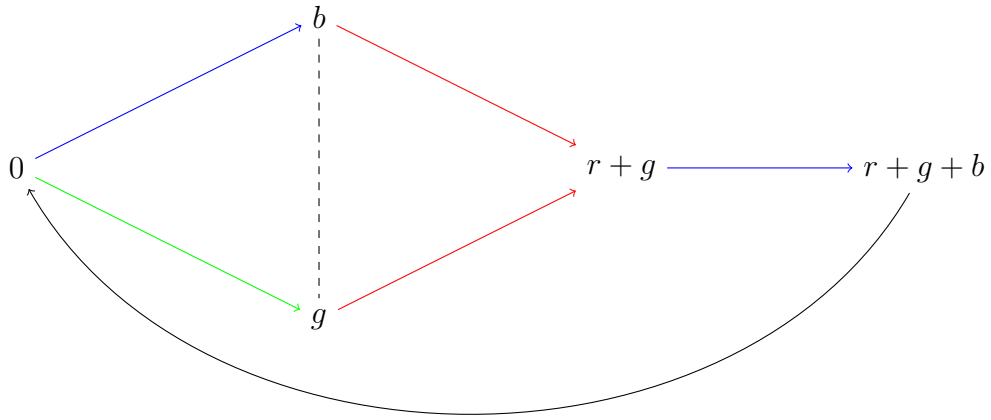



Figure 4.6: There is no way to add an edge to the depicted subgraph without creating a bad subgraph.

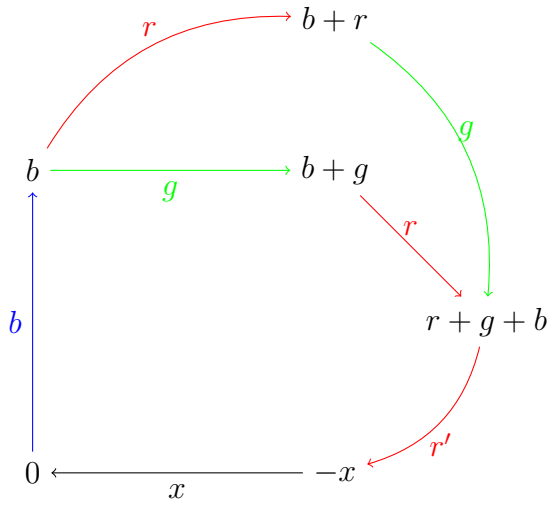



Figure 4.7: If the graph $G$ contains a 5-cycle, then we may assume it contains the subgraph depicted.

By Proposition 4.4.5, $p/4 < |A^{(v+w)}| < p/3$, and by Cauchy-Davenport

$$|3A_0^{(v+w)} + A^{(v+w)}| \geqslant 4|A^{(v+w)}| \geqslant p.$$

Thus, $3A_0^{(v+w)} + A^{(v+w)}$ is the whole of $\mathbb{Z}/p\mathbb{Z}$. Consequently, for any $g \in A_{\text{green}}$, we can find $a_1, \ldots, a_4 \in A_0^{(v+w)}$ not all zero such that $a_1 + a_2 + a_3 + a_4 = -g$. Note that $-g \notin A^{(v+w)} = A_{\text{red}} \cup A_{\text{blue}}$, so at least two of the $a_i$ are nonzero. Exactly two nonzero $a_i$ corresponds to a directed 3-cycle, whereas exactly three nonzero $a_i$ would correspond to a directed 4-cycle with a green edge, both of which cannot exist by the observations above. We conclude that all $a_i$ must be nonzero and that $G$ contains a (non-induced) 5-cycle $C$ with exactly one green edge.

The cycle $C$ must contain at least one red edge and at least one blue edge. Without loss of generality there are at least two red edges, so we may assume that $r = a_1 \in A_{\text{red}}$, $r' = a_2 \in A_{\text{red}}$ and $b = a_3 \in A_{\text{blue}}$. Set $x = a_4 \in A_{\text{red}} \cup A_{\text{blue}}$.

We will now consider subgraphs induced on (subsets of) $\{0, b, b+g, b+r, b+g+r, -x\}$ as in Figure 4.7.

First, consider $H = G[\{0, b, b+g, r+g+b, -x\}]$. Since $G$ has no induced 5-cycles, there must be another edge present. As shown in Figure 4.8, there are only two possible edges that can be added without creating a 3-cycle or a 4-cycle containing a green edge, from 0 to $b+g$ and from $b$ to $r+g+b$. Since $\{0, b, b+g, r+g+b, -x\}$ is an odd-sized set, $H$ will be bad unless there is a vertex for which the indegree has the same parity in every colour, and also a vertex for which the outdegree has the same parity in every colour.

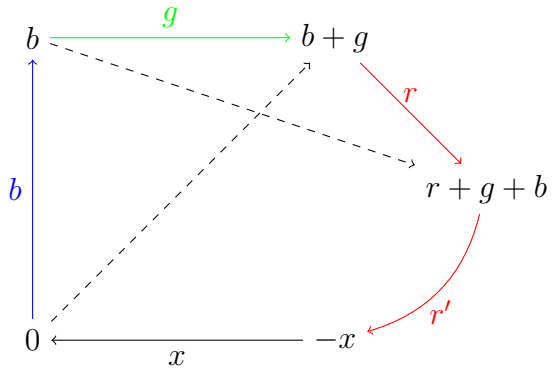



Figure 4.8: The only edges that can be added to the subgraph without creating a rainbow 4-cycle or a directed triangle are depicted with dotted lines.

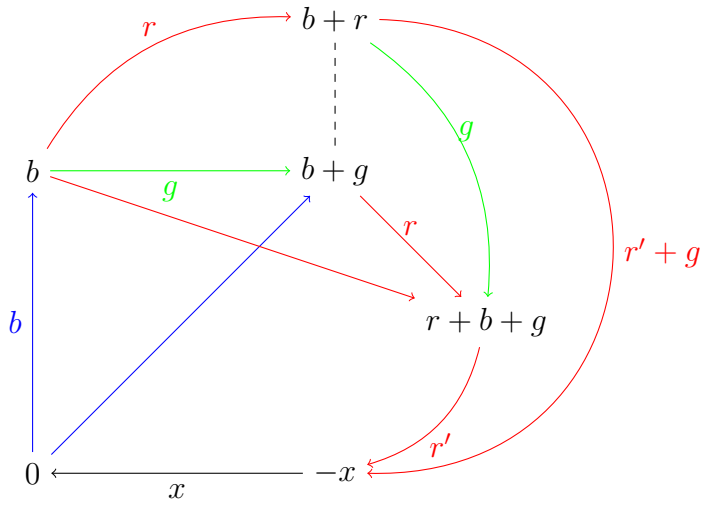



Figure 4.9: The dotted line indicates the only place where we might be able to add an edge to the subgraph.

As we cannot add green edges without creating a 4-cycle with a green edge, $H$ will be bad because of indegree unless we have a red edge from $b$ to $b+g+r$. Similarly, $H$ will be bad because of outdegree unless we have a blue edge from 0 to $b+g$.

A similar analysis on $\{0, b, b+r, r+g+b, -x\}$ shows us that there is a red edge from $b+r$ to $-x$, and so the only unknown edge is a possible edge between $b+g$ and $b+r$, as illustrated by Figure 4.9. We make the following two claims:

1. There must be either a red edge from $b+g$ to $b+r$ or a red edge from $b+r$ to $b+g$.

2. There must be either a red edge from $b+g$ to $b+r$ or a blue edge from $b+r$ to $b+g$.

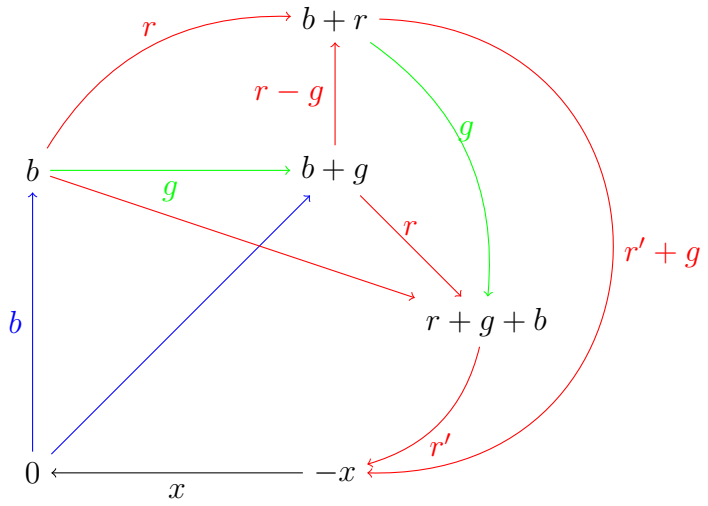



Figure 4.10: Starting from the 5-cycle depicted in Figure 4.8, we show $G$ either contains a bad subgraph or it contains the subgraph depicted in the figure.

To see the first claim, consider $H = G[\{0, b, b+g, b+r, -x\}]$. The only way that $H$ can have a vertex all of whose indegrees have the same parity is if there is a red edge from $b+g$ to $b+r$ or a red edge from $b+r$ to $b+g$. Similarly, the second claim follows by considering the outdegrees of $G[\{0, b+g, b+r, b+g+r, -x\}]$.

The only way both 1 and 2 can hold is if there is a red edge from $b+g$ to $b+r$. In particular, $r - g \in A_{\text{red}}$.

Thus, given a 5-cycle $b+r+g+r'+x = 0$, we get that $r-g$ and $r'+g$ are red, and $b+(r-g)+g+(r'+g)+x = 0$ is also a 5-cycle. We may iterate this argument to see that each element of the sequence $r, r-g, r-2g, \ldots$ will be red. However, $p$ is prime, which means that the sequence will hit every element of $\mathbb{Z}/p\mathbb{Z}$. This yields the required contradiction, and so $G$ must have a bad subgraph. □

Lemmas 4.5.4, 4.5.6 and 4.5.7 combine to show that regardless of the girths of $v$ and $w$, $G$ must have a bad subgraph, and so $v$ and $w$ cannot work together with $\mathbf{e}$. This completes the proof of Theorem 4.1.1.

## 4.6 Generalisation to arbitrary finite fields

We will now investigate $h_q(n)$, the maximal codimension of a cyclically covering subspace in $\mathbb{F}_q^n$, for $q$ a prime power. We consider this problem in the orthogonal complement in an analogous way to the case $q = 2$: we say that a vector $v \in \mathbb{F}_q^n$ *works* if for every $x \in \mathbb{F}_q^n$ there is a $k \in \{0, \ldots, n-1\}$ such that $v \cdot \sigma^k x = 0$, and that the

vectors $v^{(1)}, v^{(2)}, \cdots, v^{(m)}$ *work together* if for every $x \in \mathbb{F}_q^n$ there exists a $k$ such that

$$v^{(1)} \cdot \sigma^k x = v^{(2)} \cdot \sigma^k x = \cdots = v^{(m)} \cdot \sigma^k x = 0.$$

As before, $h_q(n)$ is given by the largest $m$ for which there exist $v^{(1)}, v^{(2)}, \ldots, v^{(m)}$ that are linearly independent and work together.

The following result is an analogue of Theorem 4.2.1, and can be proven by replacing 2 with $q$ in the proof of Theorem 4.2.1.

**Theorem 4.6.1.** *Let $q$ be a prime power. For all $m, n \in \mathbb{N}$, $h_q(mn) \geqslant h_q(m) + h_q(n)$.*

Theorem 4.2.3 generalises as follows.

**Theorem 4.6.2.** *Let $p$ be a prime, and let $q$ be a power of $p$. Then $h_q(pn) \leqslant p h_q(n)$.*

To prove Theorem 4.6.2, we use the following lemma.

**Lemma 4.6.3.** *Suppose that $f$ is a polynomial of degree at most $p-1$ over $\mathbb{F}_q$. Then*

$$\sum_{x=0}^{p-1} f(x) = \begin{cases} 0 & \deg f < p-1 \\ -c_{p-1} & \deg f = p-1, \end{cases} \tag{4.6.1}$$

*where $c_{p-1}$ refers to the leading coefficient of $f$ when $f$ has degree $p-1$.*

*Proof.* Lemma 4.6.3 will follow from the special case of monomials. If $f$ is constant, then clearly $\sum_x f(x) = pf(0) = 0$. If $f(x) = x^r$ for some $0 < r < p-1$, then let $\omega \in \mathbb{Z}/p\mathbb{Z}$ be a primitive root. Then,

$$\sum_x x^r = \sum_x (\omega x)^r = \omega^r \sum_x x^r,$$

where the first equality follows from the fact that multiplying by $\omega$ is a bijection on $(\mathbb{Z}/p\mathbb{Z})^\times$. Thus $(\omega^r - 1)\sum_x f(x) = 0$, and so $\sum_x f(x) = 0$ as $\omega$ is a primitive root. Finally, observe that $\sum_x x^{p-1} = \sum_{x \neq 0} 1 = p-1$. □

*Proof of Theorem 4.6.2.* Given a vector $v \in \mathbb{F}_q^{pn}$, we will define its *degree* as follows. For $i \in \{0, \ldots, n-1\}$ and $v \in \mathbb{F}_q^{pn}$, let

$$\begin{aligned} \pi_i(v) : \mathbb{F}_p &\to \mathbb{F}_q \\ j &\mapsto v_{nj+i}. \end{aligned}$$

Any function taking $\mathbb{F}_p$ to $\mathbb{F}_q$ can be written uniquely as the restriction to $\mathbb{F}_p$ of a polynomial of degree at most $p-1$ over $\mathbb{F}_q$. Let $d_i$ be the degree of $\pi_i(v)$. The degree $d$ of $v$ is defined as $d = \max_i d_i$.

Suppose that $V \subseteq \mathbb{F}_q^{pn}$ is the vector space spanned by $h_q(np)$ vectors that are linearly independent and work together. For $d \in \{0, \ldots, p-1\}$, let $V_d \leqslant V$ be the subspace consisting of all vectors of degree at most $d$. Since every vector has degree at most $p-1$, we immediately get that $V_{p-1} = V$.

Then, Theorem 4.6.2 will follow from the following claim.

**Claim.** *If $V_{-1}$ is defined to be the trivial subspace, then for each $0 \leqslant d \leqslant p-1$, $\dim V_d \leqslant \dim V_{d-1} + h_q(n)$.*

*Proof of Claim.* Suppose that $v^{(1)}, \ldots, v^{(r)}$ are vectors in $V_d$ that extend a basis of $V_{d-1}$ to a basis of $V_d$; thus, $r = \dim V_d - \dim V_{d-1}$. We must prove that $r \leqslant h_q(n)$.

For $t \in [r]$, define $w^{(t)} \in \mathbb{F}_q^n$ such that $w_i^{(t)}$ is the coefficient of $X^d$ in the polynomial corresponding to $\pi_i(v^{(t)})$. We claim that $w^{(1)}, \ldots, w^{(r)}$ are linearly independent and work together, proving $r \leqslant h_q(n)$.

We first show they work together. Let $x \in \mathbb{F}_q^n$. We must prove that there is a shift $k$ such that $w^{(t)} \cdot \sigma^k x = 0$ for each $t \in [r]$.

Define the vector $y \in \mathbb{F}_q^{pn}$ by $y_{nj+i} = j^{p-1-d} x_i$, for $i \in \{0, \ldots, n-1\}$ and $j \in \{0, \ldots, p-1\}$. Since the $v^{(t)}$ work together, there must be a shift $k$ with the property that $v^{(t)} \cdot \sigma^k y = 0$ for all $t \in [r]$. By definition,

$$0 = v^{(t)} \cdot \sigma^k y = \sum_{i=0}^{n-1} \sum_{j=0}^{p-1} v_{nj+i}^{(t)} y_{nj+i-k}.$$

Now, we will prove that

$$\sum_{j=0}^{p-1} v_{nj+i}^{(t)} y_{nj+i-k} = -w_i^{(t)} x_{i-k} \tag{4.6.2}$$

for each $t \in [r]$, where the index of $x_{i-k}$ is taken modulo $n$.

To prove this, note that by definition of $w^{(t)}$, we may write

$$v_{nj+i}^{(t)} = (\pi_i(v^{(t)}))(j) = w_i^{(t)} j^d + f(j)$$

where $f$ is some polynomial of degree at most $d-1$. There is a unique $\lambda \in \mathbb{Z}/p\mathbb{Z}$ with the property that, for each $j \in \{0, \ldots, p-1\}$,

$$nj + i - k = n(j + \lambda) + \ell,$$

where $\ell \in \{0, \ldots, n-1\}$ satisfies $\ell \equiv i - k \mod n$. This gives that $y_{nj+i-k} = x_{i-k}(j+\lambda)^{p-1-d}$, and thus

$$\begin{aligned}\sum_{j=0}^{p-1} v_{nj+i}^{(t)} y_{nj+i-k} &= \sum_{j=0}^{p-1} \left( w_i^{(t)} j^d + f(j) \right) x_{i-k}(j+\lambda)^{p-1-d} \\ &= -w_i^{(t)} x_{i-k}\end{aligned}$$

by Lemma 4.6.3.

It follows that

$$0 = -v^{(t)} \cdot \sigma^k y = \sum_{i=0}^{n-1} w_i^{(t)} x_{i-k} = w^{(t)} \cdot \sigma^k x,$$

for each $t \in [r]$. This proves that the $w^{(t)}$ work together.

It remains to show that the $w^{(t)}$ are linearly independent. Suppose otherwise; then, there must exist a linear relation

$$\sum_{t=1}^{r} \lambda_t w^{(t)} = 0$$

for some $\lambda_t \in \mathbb{F}_q$ not all zero.

By definition, this means that each $\pi_i(\sum_{t=1}^r \lambda_t v^{(t)}) : \mathbb{F}_p \to \mathbb{F}_q$ has coefficient of $X^d$ equal to zero, and thus has degree at most $d-1$. In other words, $\sum_{t=1}^r \lambda_t v^{(t)} \in V_{d-1}$, contradicting the assertion that the $v^{(t)}$ extend a basis of $V_{d-1}$. $\Diamond$

Applying the claim, we conclude that

$$h_q(pn) = \dim V = \sum_{d=0}^{p-1} (\dim V_d - \dim V_{d-1}) \leqslant p h_q(n). \qquad \square$$

For $\ell < p$, note that $h_q(\ell) \leqslant \log_q(\ell) < 1$. This gives us the following corollary of Theorem 4.6.2.

**Corollary 4.6.4.** *Let $p$ be prime, and let $q$ be a power of $p$. Then $h_q(\ell p^d) = 0$ for any $\ell < q$.*

*Remark.* In Theorem 4.6.2, we require that the characteristic of the field $\mathbb{F}_q$ is the same $p$ as in the bound. For example, $h_3(2n) \leqslant 2h_3(n)$ fails for $n = 4$. Indeed, observe that $h_3(4) = 0$ (as noted in [11]), whereas $h_3(8) = 1$ follows from [11, Theorem 5].

For $q > 2$, we have the following stronger version of Theorem 4.1.1.

**Theorem 4.6.5.** *Suppose that $q$ is an odd prime, and $p > q$ is a prime with $q$ as a primitive root. Then $h_q(p) = 0$; in other words, there are* no *nonzero vectors $v \in \mathbb{F}_q^p$ that work.*

*Remark.* It is known by a result of Heath-Brown [34] that for all but at most two primes $q$, there are infinitely many primes $p$ such that $q$ is a primitive root of $p$. In particular, we know unconditionally that for all but two primes $q$, $h_q(n)$ does not tend to infinity among $n$ coprime to $q$. It is widely believed that there are no primes $q$ that are primitive roots for only finitely many $p$; if this were true, we would know that $h_q(n)$ does not tend to infinity for any $q$ (among $n$ coprime to $q$).

*Proof.* Suppose that $v \in \mathbb{F}_q^p$ is nonzero. We consider the polynomial $f_v \in \mathbb{F}_p[X]/(X^p - 1)$ which corresponds to a vector $v$, after making the natural modifications to the definition given in Section 4.3. An analogue of Proposition 4.3.1 holds over $\mathbb{F}_q$, with the same proof. Thus, $v$ works if and only if, for every $x \in \mathbb{F}_q^p$, there is a $k$ such that the coefficient of $X^k$ in $f_v f_x$ is 0. This holds if and only if $f_v f_x$ is never equal to a polynomial with no coefficient equal to zero. In other words, $v$ fails to work if and only if $f_v$ is a factor of a polynomial with no zero coefficients.

Since $q$ is a primitive root for the prime $p$, the analogue of Lemma 4.3.2 tells us that $X^p - 1$ factors into irreducible polynomials as

$$X^p - 1 = (X - 1)(1 + X + \cdots + X^{p-1}).$$

Then we can write

$$\frac{\mathbb{F}_q[X]}{(X^p - 1)} = \frac{\mathbb{F}_q[X]}{(X - 1)} \oplus \frac{\mathbb{F}_q[X]}{(1 + X + X^2 + \cdots + X^{p-1})}.$$

Suppose $f_v$ is nonzero in $\mathbb{F}_q[X]/(X^p - 1)$. Write $a \in \mathbb{F}_q[X]/(X - 1)$ for $f_v \mod X - 1$ and $b \in \mathbb{F}_q[X]/(1 + X + \cdots + X^{p-1})$ for $f_v \mod 1 + X + \cdots + X^{p-1}$. The assertion that $f_v$ is nonzero is then equivalent to the assertion that either $a$ or $b$ is nonzero.

Suppose first that $a$ is nonzero. Since $\mathbb{F}_q[X]/(X - 1)$ is a field, there is an inverse $a^{-1}$ for $a$. Let $c = 1 + X + \cdots + X^{p-1} \mod X - 1$, and let $g \in \mathbb{F}_q[X]/(X^p - 1)$ denote the polynomial that is $ca^{-1} \mod X - 1$ and $0 \mod 1 + X + \cdots + X^{p-1}$. Thus, $f_v g$ is $c \mod X - 1$ and $0 \mod 1 + X + \cdots + X^{p-1}$, and so is equal to $1 + X + \cdots + X^{p-1} \mod X^p - 1$. In particular, $f_v$ is a factor of $1 + X + \cdots + X^{p-1}$.

On the other hand, suppose that $a$ is zero. Then $b$ must be nonzero, and a similar argument shows that $f_v$ is a factor of $X - 1$. Thus, any nonzero polynomial in $\mathbb{F}_q[X]/(X^p - 1)$ is a factor of either $X - 1$ or $1 + \cdots + X^{p-1}$.

Now, $1 + \cdots + X^{p-1}$ itself has no coefficient equal to zero. Also, $X - 1$ is a factor of

$$(X-1)(X^{p-2} + X^{p-4} + \cdots + X - 1) = 1 - 2X + X^2 + \sum_{i=3}^{p-1} (-X)^i,$$

which also has no coefficient equal to zero.

Thus, any nonzero polynomial is a factor of a polynomial with no coefficient equal to zero, and so any nonzero vector does not work. □

# Chapter 5

# Functions of large additive energy supported on a Hamming Sphere

## 5.1 Introduction

For a function $f : \mathbb{F}_2^n \to \mathbb{R}$, we define its Gowers $U_2$ norm to be

$$||f||_{U_2} = \left( \mathbb{E}_{\substack{a_1,a_2,a_3,a_4 \in \mathbb{F}_2^n \\ a_1+a_2=a_3+a_4}} f(a_1)f(a_2)f(a_3)f(a_4) \right)^{1/4}. \tag{5.1.1}$$

This is also called the *additive energy* of $f$, and agrees with the usual notion of energy for sets (up to a scaling) in the sense that $||\mathbb{1}_A||_{U_2}^4 = \frac{E(A)}{N^3}$, where $N = 2^n$ is the size of $\mathbb{F}_2^n$. Similarly, we will define the $\ell^2$ norm to be

$$||f||_2 = \left( \mathbb{E}_{a \in \mathbb{F}_2^n} f(a)^2 \right)^{1/2}.$$

For a set $A \subseteq \mathbb{F}_2^n$, define $\mu(A)$ by

$$\mu(A) = \max_{\substack{f:\mathbb{F}_2^n \to \mathbb{R} \\ \text{supp} f \subseteq A}} \frac{||f||_{U_2}^4}{||f||_2^4}, \tag{5.1.2}$$

where $\text{supp}(f)$ denotes the support of $f$.

Let the Hamming Sphere $S(n,k) \subseteq \mathbb{F}_2^n$ consist of those vectors of weight $k$; in other words, $S(n,k)$ consists of those vectors with exactly $k$ ones. A natural lower bound for $\mu(A)$ would be to take $f$ to be constant on $A$. This gives $\mu(A) \geqslant \frac{1}{N}\frac{E(A)}{|A|^2}$. In [39], Kirshner and Samorodnitsky conjectured that this is optimal:

**Conjecture** (Conjecture 1.9 from [39])**.** *Let $A = S(n,k)$. Then, $\mu(A) = \frac{1}{N}\frac{E(A)}{|A|^2}$.*

*In other words, the ratio $\frac{||f||_{U_2}^4}{||f||_2^4}$ achieves its maximum when $f$ is constant.*

The purpose of this section is to establish this conjecture.

**Theorem 5.1.1.** *Let $A = S(n,k)$. Then, $\mu(A) = \frac{1}{N}\frac{E(A)}{|A|^2}$.*

*Moreover, if $2 \leqslant k \leqslant n-2$, then any function which maximises $\frac{||f||_{U_2}^4}{||f||_2^4}$ must have $|f|$ constant.*

*Remark.* Kirshner and Samorodnitsky define $\mu(A)$ as the maximal value of $\frac{||f||_4^4}{||f||_2^4}$ among functions whose Fourier transform is supported on $A$. However, it can easily be seen that these two formulations are equivalent (up to normalisation) by taking a Fourier transform, and using Parseval's identity and the relation that $||f||_4^4 = ||\hat{f}||_{U_2}^4$.

## 5.2 Proof of Theorem 5.1.1

Throughout the proof, let $e_1, \ldots, e_n$ denote the standard basis for $\mathbb{F}_2^n$, so that any element of $\mathbb{F}_2^n$ may be written $\sum_i \varepsilon_i e_i$, where $\varepsilon_i \in \mathbb{F}_2$. If $v, w \in \mathbb{F}_2^n$, let $\langle v, w \rangle$ denote the standard inner product of $v$ and $w$. In other words,

$$\left\langle \sum_i \varepsilon_i^{(1)} e_i, \sum_i \varepsilon_i^{(2)} e_i \right\rangle = \sum_i \varepsilon_i^{(1)} \varepsilon_i^{(2)}.$$

The idea behind our approach is to define an operation on functions $f$, which we call a compression, which increases the value of $||f||_{U^2}$ without affecting $||f||_2$. In particular, there must be a function maximising the ratio $\frac{||f||_{U_2}^4}{||f||_2^4}$ which is invariant under compressions. This idea is loosely inspired by the use of compression techniques in extremal set theory, and particularly the additive combinatorial applications as in [28]. However, the actual compressions we consider are different.

**Definition 5.2.1.** For a function $f : \mathbb{F}_2^n \to \mathbb{R}$ and $i < j \leqslant n$, define the $i,j$ compression $f^{(ij)}$ as follows:

$$f^{(ij)}(x) = \begin{cases} f(x) & \langle x, e_i + e_j \rangle = 0 \\ \sqrt{\frac{f(x)^2 + f(x+e_i+e_j)^2}{2}} & \text{otherwise.} \end{cases}$$

In other words, let $\pi_{ij} : \mathbb{F}_2^n \to \mathbb{F}_2^{(n-2)}$ denote the projection given by ignoring the coefficients of $e_i$ and $e_j$. Then, $f^{(ij)}(x)$ is the $\ell^2$-average of $f$ over elements of the coset of $\ker \pi_{ij}$ containing $x$, which have the same Hamming weight as $x$.

The proof of Theorem 5.1.1 relies on the following lemma about compressions.

**Lemma 5.2.2.** *Let $A = S(n,k)$, and suppose that $f$ is supported on $A$.*

1. *$f^{(ij)}$ is also supported on $A$.*

*2.* $||f^{(ij)}||_2 = ||f||_2$.

*3.* $||f^{(ij)}||_{U_2} \geqslant ||f||_{U_2}$.

*4. If equality holds above, then it must be the case that* $|f| = |f^{(ij)}|$.

*Proof.* The proofs of (1) and (2) follow immediately from Definition 5.2.1.

For (3), observe that we may rewrite (5.1.1) as follows.

$$||f||_{U_2}^4 = \frac{1}{4^3}\mathbb{E}_{\substack{b_1,b_2,b_3,b_4\in\pi_{ij}(\mathbb{F}_2^n)\\ b_1+b_2=b_3+b_4}}\left(\sum_{\substack{a_t\in\pi_{ij}^{-1}(b_i)\cap A\\ a_1+a_2=a_3+a_4}} f(a_1)f(a_2)f(a_3)f(a_4)\right), \tag{5.2.1}$$

where the outer expectation is over cosets of $\ker\pi_{ij}$, and the factor of $\frac{1}{4^3}$ comes from the fact that we have renormalised the inner expectation to be a summation. Our strategy will be to prove that each bracketed term on the right hand side of (5.2.1) does not decrease when we pass from $f$ to $f^{(ij)}$.

Observe that, if $f$ is supported on $A$, then the outer expectation of (5.2.1) may be restricted to terms such that each $b_t$ has Hamming weight either $k, k-1$ or $k-2$, and the size of $\pi_{ij}^{-1}(b_t)\cap A$ depends on whether $b_t$ has weight $k-1$ or not. Thus, we split naturally into three cases.

*Case 1: None of* $b_1, b_2, b_3$ *or* $b_4$ *has Hamming weight* $k-1$. In this case, the bracketed term is a sum over exactly one term, and is unchanged as we pass from $f$ to $f^{(ij)}$.

*Case 2: Exactly two of* $b_1, b_2, b_3$ *and* $b_4$ *have Hamming weight* $k-1$. Without loss of generality, it is $b_1$ and $b_2$ which have Hamming weight $k-1$. Then, there are two possibilities for the bracketed term, depending on how many of $b_3$ and $b_4$ have weight $k-2$. If neither or both of them do, then we may write the bracketed term as

$$f(b_1+e_i)f(b_2+e_i)f(a_3)f(a_4) + f(b_1+e_j)f(b_2+e_j)f(a_3)f(a_4),$$

where $a_3$ denotes the unique element of $\pi_{ij}^{-1}(b_3)\cap A$ (and likewise for $a_4$). The conclusion then follows from the assertion that

$$f(b_1+e_i)f(b_2+e_i) + f(b_1+e_j)f(b_2+e_j) \leqslant 2f^{(ij)}(b_1+e_i)f^{(ij)}(b_2+e_i),$$

which is a consequence of the Cauchy-Schwarz inequality. A similar argument applies if exactly one of $b_3$ and $b_4$ have weight $k-2$.

*Case 3: All four of $b_1, b_2, b_3$ and $b_4$ have Hamming weight $k-1$.* In this case, the bracketed term is now a sum of eight terms. One of the terms is

$$f(a_1)f(a_2)f(a_3)f(a_4) = f(b_1+e_i)f(b_2+e_i)f(b_3+e_i)f(b_4+e_i),$$

and the others can be obtained by replacing two or four of the $e_i$ with $e_j$.

Group the terms into four pairs, according to the values of $a_3$ and $a_4$. If $a_3 = b_3 + e_1$ and $a_4 = b_4 + e_1$, for example, then we have

$$\begin{aligned} f(b_1+e_i)f(b_2+e_i)f(a_3)f(a_4) + f(b_1+e_j)f(b_2+e_j)f(a_3)f(a_4) \\ \leqslant f^{(ij)}(a_1)f^{(ij)}(a_2)f(a_3)f(a_4), \end{aligned}$$

as in case 2. The conclusion then follows from the fact that

$$f(b_3+e_i) + f(b_3+e_j) \leqslant 2f^{(ij)}(a_3),$$

which follows from Cauchy-Schwarz.

Finally, it remains to prove (4). Suppose that $|f| \neq |f^{(ij)}|$. In other words, there must be some vector $v$ of weight $k-1$ such that $f(v+e_i) \neq f(v+e_j)$. Then, consider the term of (5.2.1) coming from $b_1 = b_2 = b_3 = b_4 = v$. This term may be rewritten as

$$f(v+e_i)^4 + 6f(v+e_i)^2 f(v+e_j)^2 + f(v+e_j)^4.$$

The corresponding term for $f^{(ij)}$ is $8f^{(ij)}(v+e_j)^4$ in view of the fact that $f^{(ij)}(v+e_i) = f^{(ij)}(v+e_j)$, and this is strictly larger for $f^{(ij)}$ than it is for $f$ by the AM-GM inequality. □

We can now complete the proof of Theorem 5.1.1. By compactness, there must exist some function $f$ achieving the maximal value of $||f||_{U_2}$, for fixed $||f||_2$. Suppose that this maximal value is achieved for a function $f$ for which $|f|$ is not constant.

Consider the Hamming Sphere as a graph, where we join two elements $v$ and $w$ with an edge if and only if $w = v + e_i + e_j$ for some $i$ and $j$. Then, the Hamming Sphere is connected. Thus, there must be two adjacent elements $v$ and $w$ for which $|f(v)| \neq |f(w)|$.

Thus, if $w = v + e_i + e_j$, then $|f| \neq |f^{(ij)}|$, and so Lemma 5.2.2 (4) tells us that $||f||_{U_2} < ||f^{(ij)}||_{U_2}$, contradicting the maximality of $||f||_{U_2}$.

Therefore, $|f|$ must be constant. Observe that by the definition of $U_2$ norm, the $U_2$ norm of $f$ is bounded above by the $U_2$ norm of $|f|$, proving that the constant function indeed maximises $||f||_{U_2}$ among functions with $||f||_2$ fixed. □

*Remark.* In fact, by examining the proof of Lemma 5.2.2, we may classify exactly which functions $f$ with $|f|$ constant maximise $||f||_{U_2}/||f||_2$. We will prove that any function maximising this ratio must be of the form $f(v) = C(-1)^{v\cdot w}$, where $w$ is a fixed vector. In particular, if $k = 1$ or $n-1$ then any function with $|f|$ constant maximises the ratio.

Suppose that $v$ and $w$ are any vectors of weight $k-1$ with no $e_i$ or $e_j$ component, and let $e$ be any basis vector that is not $e_i$ or $e_j$. For the term in (5.2.1) coming from $b_1 = v, b_2 = w, b_3 = v+e, b_4 = w+e$, the bracketed term becomes

$$f(v+e_i)f(w+e_i)f(v+e)f(w+e) + f(v+e_j)f(w+e_j)f(v+e)f(w+e).$$

For this not to increase when we pass to $f^{(ij)}$, it must be the case that we have $f(v+e_i)f(w+e_i) = f(v+e_j)f(w+e_j)$ (noting that all of the factors are nonzero as $|f|$ is constant). In other words, the signs of $f(v+e_i)$ and $f(v+e_j)$ are the same if and only if the signs of $f(w+e_i)$ and $f(w+e_j)$ are the same.

Consider the graph on vertex set $[n]$, with an edge between $i$ and $j$ if the signs of $f(v+e_i)$ and $f(v+e_j)$ are different for all $v$. Such a graph must be a complete bipartite graph (or an empty graph, which may be viewed as a complete bipartite graph with one of the parts being empty). Indeed, for any $i, j$ and $k$, the number of edges among $ij, jk$ and $ik$ that are present must be even, and an easy induction shows that such a graph must be complete bipartite.

Accordingly, let $w$ be the sum of the $e_i$ with $i$ taken from one of the parts. It is easy to see that the sign of $f(v)(-1)^{v\cdot w}$ must remain constant, and thus $f(v) = C(-1)^{v\cdot w}$ as claimed.

# Chapter 6

# A counterexample to a strong variant of the Polynomial Freiman-Ruzsa conjecture

## 6.1 Introduction

Let $A \subseteq \mathbb{F}_p^n$ be a subset, with $|A + A| \leqslant K|A|$. The Polynomial Freiman-Ruzsa conjecture, attributed to Marton, asserts that $A$ can be covered by at most $K^{O(1)}$ translates of a subspace of size at most $K^{O(1)}|A|$.

An equivalent form of the Polynomial Freiman-Ruzsa conjecture may be stated as follows [24]:

**Conjecture 6.1.1.** *Suppose that $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$ is a function such that $\phi(x+y) - \phi(x) - \phi(y)$ takes values in some set $S$.*

*Then there is a linear map $\tilde{\phi} : \mathbb{F}_p^n \to \mathbb{F}_p^N$ with the property that $\phi - \tilde{\phi}$ takes at most $|S|^{O(1)}$ values.*

It is tempting to hope that there might be a natural combinatorial relationship between the set $S$ and the image of $\phi - \tilde{\phi}$. The following is a natural strengthening of Conjecture 6.1.1, which has been considered elsewhere (see [19], [53, Section 1.17]).

**Conjecture 6.1.2.** *There is a constant $t$ such that the following holds. Suppose that $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$ is a function such that $\phi(x+y) - \phi(x) - \phi(y)$ takes values in some set $S$.*

*Then there is a linear map $\tilde{\phi} : \mathbb{F}_p^n \to \mathbb{F}_p^N$ with the property that $\phi - \tilde{\phi}$ takes values in $tS$.*

In particular, in [19] Farah proves Conjecture 6.1.2 when $S$ is a Hamming ball.

Green and Tao give a counterexample in [53, Section 1.17]. However, their example is based on quadratic forms, and only works in the case that $p = 2$. In [19, Theorem 3.3], Farah proposes a different example which refutes Conjecture 6.1.2, and works for any value of $p$.

In this section, we offer an alternative approach to that of Farah to refute Conjecture 6.1.2. Our construction defines the same function $\phi$ as Farah, but provides a somewhat different argument to prove that $\phi - \tilde{\phi}$ does not take values in $tS$. In particular, our approach gives linear bounds on how large we need to take $n$ to refute the conjecture for a particular value of $t$; namely, we show that we may take $n = 12t + 7$. By contrast, Farah's approach provides bounds on $n$ which are doubly exponential in $t$, as the argument relies on Ramsey's theorem.

Note that our construction is optimal, in the sense that there cannot be any counterexample to Conjecture 6.1.2 with a bound better than linear.

**Proposition 6.1.3.** *If $t \geqslant (p-1)n$, then Conjecture 6.1.2 holds.*

*Proof.* Given $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$, construct $\tilde{\phi}$ as follows. Let $e_1, \ldots, e_n$ form a basis for $\mathbb{F}_p^n$ and say that if $x = x_1e_1 + \cdots + x_ne_n$ for scalars $x_i \in \mathbb{F}_p$, then $\tilde{\phi}(x) = x_1\phi(e_1) + \cdots + x_n\phi(e_n)$.

Now, if $x = x_1e_1 + \cdots + x_ne_n$ with $x_1 + \cdots + x_n = k$ (here considering the $x_i$ as elements of the interval $[0, p-1] \subseteq \mathbb{Z}$), we will prove that $\phi(x) - \tilde{\phi}(x)$ is the sum of at most $k$ elements of $S$ via an induction on $k$. In particular, this will complete the proof of Proposition 6.1.3.

If $k = 1$ then $\phi(x) - \tilde{\phi}(x) = 0$. For the inductive step, we simply use the relation

$$\phi(x) - \tilde{\phi}(x) = \phi(x - e_i) - \tilde{\phi}(x - e_i) + (\phi(x) - \phi(x - e_i) - \phi(e_i))$$

which follows from the definition of $\tilde{\phi}$. If we choose $i$ so that $x_i \neq 0$, then the inductive hypothesis tells us that $\phi(x - e_i) - \tilde{\phi}(x - e_i)$ may be written as the sum of $k - 1$ elements of $S$, which completes the proof. □

## 6.2 The construction

The construction goes as follows.

**Theorem 6.2.1.** *Given $t$, suppose that $n \geqslant 12t + 7$. View $\mathbb{F}_p^N$ as the space of all (not necessarily linear) maps from $\mathbb{F}_p^{\mathbb{F}_p^n}$ to $\mathbb{F}_p$ (in particular, $N = p^{p^n}$).*

*Let $\phi : \mathbb{F}_p^n \to \mathbb{F}_p^N$ be defined as follows. If $v$ is an element of $\mathbb{F}_p^n$, then define $\phi(v) : \mathbb{F}_p^{\mathbb{F}_p^n} \to \mathbb{F}_p$ by*

$$\phi(v) : f \to f(v),$$

*where $f : \mathbb{F}_p^n \to \mathbb{F}_p$.*

*Let $S \subseteq \mathbb{F}_p^N$ denote the set of values taken by $\phi(x+y) - \phi(x) - \phi(y)$. Then, there is no linear map $\tilde{\phi}$ such that $\phi - \tilde{\phi}$ takes values in $tS$.*

*Proof.* Suppose that there does exist $\tilde{\phi}$ such that $\phi - \tilde{\phi}$ takes values in $tS$.

An element of $S$ may be written $\phi(a+b) - \phi(a) - \phi(b)$ for some $a, b \in \mathbb{F}_p^n$. Thus, for each $x \in \mathbb{F}_p^n$, there are pairs $(a_i^{(x)}, b_i^{(x)})$ for $1 \leqslant i \leqslant t$ with the property that

$$(\phi(x) - \tilde{\phi}(x))(f) = \sum_{i=1}^{t} f(a_i^{(x)} + b_i^{(x)}) - f(a_i^{(x)}) - f(b_i^{(x)}) \tag{6.2.1}$$

for every map $f : \mathbb{F}_p^n \to \mathbb{F}_p$.

For each $x \in \mathbb{F}_p^n$, define $V_x = \mathrm{Span}(x, a_1^{(x)}, b_1^{(x)}, \ldots, a_t^{(x)}, b_t^{(x)})$. $V_x$ obeys the following three properties:

1. $V_x$ has dimension at most $2t+1$.
2. $x \in V_x$.
3. Suppose that $f : \mathbb{F}_p^n \to \mathbb{F}_p$ is a map such that $\phi(x)(f) \neq \tilde{\phi}(x)(f)$. Then the restriction $f|_{V_x}$ is nonlinear.

The first two properties follow trivially. To see why the third holds, suppose that $\phi(x)(f) \neq \tilde{\phi}(x)(f)$. By (6.2.1), we learn that

$$\sum_{i=1}^{t} f(a_i^{(x)} + b_i^{(x)}) - f(a_i^{(x)}) - f(b_i^{(x)}) \neq 0.$$

Thus, at least one term in the sum must be nonzero; from that term, we have that

$$f(a_i^{(x)} + b_i^{(x)}) \neq f(a_i^{(x)}) + f(b_i^{(x)}).$$

However, all three of the arguments to $f$ above are in $V_x$, so $f|_{V_x}$ must be nonlinear.

Now, in order to find a contradiction, observe that it suffices to find a pair $x, y \in \mathbb{F}_p^n$ such that

$$x + y \notin (V_x \cap V_{x+y}) + (V_y \cap V_{x+y}). \tag{6.2.2}$$

To see why, construct a function $f : \mathbb{F}_p^n \to \mathbb{F}_p$ as follows:

- Set $f|_{V_x} = 0$ and $f|_{V_y} = 0$.
- Set $f(x+y) = 1$, and extend to a linear function on $V_{x+y}$. This is possible because of our condition on $x+y$.
- Define $f$ arbitrarily on $\mathbb{F}_p^n \backslash (V_x \cup V_y \cup V_{x+y})$.

Now, we have that

$$\begin{aligned} 0 = f(x) = \phi(x)(f) = \tilde{\phi}(x)(f) \\ 0 = f(y) = \phi(y)(f) = \tilde{\phi}(y)(f) \\ 1 = f(x+y) = \phi(x+y)(f) = \tilde{\phi}(x+y)(f). \end{aligned}$$

In each line, the first equality holds by construction, the second follows from the definition of $\phi$ and the third follows from property (3) since $f|_{V_x}, f|_{V_y}$ and $f|_{V_{x+y}}$ are linear. Thus $\tilde{\phi}(x+y) - \tilde{\phi}(x) - \tilde{\phi}(y)$ is not zero when evaluated at $f$, contradicting the linearity of $\tilde{\phi}$.

Thus, we will be done if we can establish the following:

**Proposition 6.2.2.** *Suppose that for each $x \in \mathbb{F}_p^n$, we have a subspace $V_x$ of dimension at most $2t+1$. Then, provided that $n \geqslant 12t+7$, there must exist some pair $x, y$ satisfying (6.2.2).*

*Proof.* Suppose that no such pair $x, y$ exist. Then, for each $x, z$ there exist $v_{x,y} \in V_x \cap V_z$ and $w_{x,z} \in V_{z-x} \cap V_z$, with $z = v_{x,z} + w_{x,z}$.

As $x$ runs over $\mathbb{F}_p^n$, $v_{x,z}$ takes values in $V_z$. Thus, by the pigeonhole principle, there must exist some choice $v_z \in V_z$ which occurs for at least $p^{n-4t-2}$ choices of $x$.

If we define $U = \{v \in \mathbb{F}_p^n | v \in V_x \text{ for at least } p^{n-4t-2} \text{ choices of } x\}$, then we instantly learn that $v_z \in U$ for each $z \in \mathbb{F}_p^n$, since $v_z \in V_x$ for at least $p^{n-4t-2}$ choices of $x$. Similarly, $w_z \in U$, since $w_z \in V_{z-x}$. In view of the fact that $v_z + w_z = z$, we learn that $U + U$ is the whole of $\mathbb{F}_p^n$, and so $|U| \geqslant p^{n/2}$.

However, we can also give an upper bound for $|U|$. There are at most $p^{n+2t+1}$ pairs $v, x$ with $v \in V_x$, and each $v \in U$ must count at least $p^{n-4t-2}$ of those pairs. Thus, there can be at most $p^{6t+3}$ elements of $U$.

Putting this together, we have that $p^{n/2} \leqslant |U| \leqslant p^{6t+3}$, and so that $n \leqslant 12t+6$. □

This gives us the required contradiction, and so such a linear function $\tilde{\phi}$ indeed cannot exist. □

# Bibliography


[1] J. Aaronson. Maximising the number of solutions to a linear equation in a set of integers. *Bull. Lond. Math. Soc.*, 51(4):577–594, 2019.

[2] N. Alon and A. Shapira. Testing subgraphs in directed graphs. *J. Comput. System Sci.*, 69(3):353–382, 2004.

[3] W. Beckner. Inequalities in Fourier analysis. *Ann. of Math. (2)*, 102(1):159–182, 1975.

[4] F. Behrend. On sets of integers which contain no three terms in arithmetical progression. *Proc. Nat. Acad. Sci. U. S. A.*, 32:331–332, 1946.

[5] B. Bhattacharya, S. Ganguly, X. Shao, and Y. Zhao. Upper tail large deviations for arithmetic progressions in a random set. *Int. Math. Res. Not. IMRN*, 2020.

[6] J. Blasiak, T. Church, H. Cohn, J. Grochow, E. Naslund, W. Sawin, and C. Umans. On cap sets and the group-theoretic approach to matrix multiplication. *Discrete Anal.*, 2017.

[7] A. Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier (Grenoble)*, 20(fasc., fasc. 2):335–402 (1971), 1970.

[8] D. Borwein and J. Borwein. Some remarkable properties of sinc and related integrals. *Ramanujan J.*, 5(1):73–89, 2001.

[9] B. Bukh. Sums of dilates. *Combin. Probab. Comput.*, 17(5):627–639, 2008.

[10] P.J. Cameron. Problems from the Thirteenth British Combinatorial Conference. *Discrete Mathematics*, 125:407–417, 1994.

[11] P.J. Cameron, D. Ellis, and W. Raynaud. Smallest cyclically covering subspaces of $\mathbb{F}_q^n$, and lower bounds in Isbell's conjecture, 2019. arXiv:1810.03485v2.

[12] P.J. Cameron, P. Frankl, and W.M. Kantor. Intersecting families of finite sets and fixed-point-free 2-elements. *European Journal of Combinatorics*, 10(2):149–160, 1989.

[13] D. Conlon and J. Fox. Graph removal lemmas. *Surveys in combinatorics*, 1(2):3, 2013.

[14] E. Croot. The minimal number of three-term arithmetic progressions modulo a prime converges to a limit. *Canad. Math. Bull.*, 51(1):47–56, 2008.

[15] E. Croot, V. Lev, and P. Pach. Progression-free sets in $\mathbb{Z}_4^n$ are exponentially small. *Ann. of Math. (2)*, 185(1):331–337, 2017.

[16] S. Eberhard. The abelian arithmetic regularity lemma, 2016. arXiv:1606.09303.

[17] S. Eberhard, B. Green, and F. Manners. Sets of integers with no large sum-free subset. *Ann. of Math. (2)*, 180(2):621–652, 2014.

[18] J. Ellenberg and D. Gijswijt. On large subsets of $\mathbb{F}_q^n$ with no three-term arithmetic progression. *Ann. of Math. (2)*, 185(1):339–343, 2017.

[19] I. Farah. Approximate homomorphisms. II. Group homomorphisms. *Combinatorica*, 20:47–60, 2000.

[20] J. Fox and L.M. Lovász. A tight bound for Green's arithmetic triangle removal lemma in vector spaces. *Adv. Math.*, 321:287–297, 2017.

[21] R.M. Gabriel. The Rearrangement of Positive Fourier Coefficients. *Proc. London Math. Soc. (2)*, 33(1):32–51, 1931.

[22] S. Ganguly. Personal communication, 2018.

[23] L.S. Goddard. The accumulation of chance effects and the gaussian frequency distribution. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 36(257):428–433, 1945.

[24] B. Green. Finite field models in additive combinatorics. In *Surveys in combinatorics 2005*, volume 327 of *London Math. Soc. Lecture Note Ser.*, pages 1–27. Cambridge Univ. Press, Cambridge, 2005.

[25] B. Green. A Szemerédi-type regularity lemma in abelian groups, with applications. *Geom. Funct. Anal.*, 15(2):340–376, 2005.

[26] B. Green and S. Lindqvist. Monochromatic solutions to $x + y = z^2$. *Canad. J. Math.*, 71(3):579–605, 2019.

[27] B. Green and O. Sisask. On the maximal number of 3-term arithmetic progressions in subsets of $\mathbb{Z}/p\mathbb{Z}$. *Bull. Lond. Math. Soc.*, 40(6):945–955, 2008.

[28] B. Green and T. Tao. Freiman's theorem in finite fields via extremal set theory. *Combin. Probab. Comput.*, 18(3):335–355, 2009.

[29] B. Green and T. Tao. An arithmetic regularity lemma, an associated counting lemma, and applications. *An irregular mind*, pages 261–334, 2010.

[30] D.J. Grynkiewicz. *Structural additive theory*, volume 30 of *Developments in Mathematics*. Springer, Cham, 2013.

[31] Y.O. Hamidoune and Ø.J. Rødseth. An inverse theorem mod $p$. *Acta Arithmetica*, 92(3):251–262, 2000.

[32] G.H. Hardy and J.E. Littlewood. Notes on the theory of series (viii): an inequality. *Journal of the London Mathematical Society*, 1(2):105–110, 1928.

[33] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge university press, 1952.

[34] D.R. Heath-Brown. Artin's conjecture for primitive roots. *The Quarterly Journal of Mathematics*, 37(1):27–38, 1986.

[35] C. Hooley. On Artin's conjecture. *Journal für die reine und angewandte Mathematik*, 225:209–220, 1967.

[36] J. Isbell. Homogeneous games. *Math. Student*, 25:123–128, 1957.

[37] J. Isbell. Homogeneous games. ii. *Proceedings of the American Mathematical Society*, 11(2):159–161, 1960.

[38] A. Khintchine. Über dyadische Brüche. *Math. Z.*, 18(1):109–116, 1923.

[39] N. Kirshner and A. Samorodnitsky. On $\ell_4 : \ell_2$ ratio of functions with restricted Fourier support, 2018. arXiv:1801.08507.

[40] N. Kirshner and A. Samorodnitsky. A moment ratio bound for polynomials and some extremal properties of krawchouk polynomials and hamming spheres, 2019. arXiv:1909.11929.

[41] R. Kleinberg. A nearly tight upper bound on tri-colored sum-free sets in characteristic 2, 2016. arXiv:1605.08416.

[42] D. Král, O. Serra, and L. Vena. A combinatorial proof of the removal lemma for groups. *J. Combin. Theory Ser. A*, 116(4):971–978, 2009.

[43] V. Lev. On the number of solutions of a linear equation over finite sets. *J. Combin. Theory Ser. A*, 83(2):251–267, 1998.

[44] V. Lev and R. Pinchasi. Solving $a \pm b = 2c$ in elements of finite sets. *Acta Arith.*, 163(2):127–140, 2014.

[45] R. Lidl and H. Niederreiter. *Finite fields*, volume 20 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, second edition, 1997.

[46] M.B. Nathanson. Growth polynomials for additive quadruples and $(h, k)$-tuples. *Acta Math. Hungar.*, 143(1):44–57, 2014.

[47] G. Polya. Berechnung eines bestimmten Integrals. *Math. Ann.*, 74(2):204–212, 1913.

[48] I. Ruzsa. Generalized arithmetical progressions and sumsets. *Acta Math. Hungar.*, 65(4):379–388, 1994.

[49] I. Ruzsa. An analog of Freiman's theorem in groups. *Astérisque*, 258(199):323–326, 1999.

[50] I. Ruzsa and E. Szemerédi. Triple systems with no six points carrying three triangles. In *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II*, volume 18 of *Colloq. Math. Soc. János Bolyai*, pages 939–945. North-Holland, Amsterdam-New York, 1978.

[51] O. Serra and G. Zémor. Large sets with small doubling modulo $p$ are well covered by an arithmetic progression. *Ann. Inst. Fourier (Grenoble)*, 59(5):2043–2060, 2009.

[52] G. Shakan. Sum of many dilates. *Combin. Probab. Comput.*, 25(3):460–469, 2016.

[53] T. Tao. *Poincaré's legacies, pages from year two of a mathematical blog. Part I.* American Mathematical Society, 2009.

[54] T. Tao and V.H. Vu. *Additive combinatorics*, volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.