On the complexity of finding and counting solution-free sets of integers

Treglown, Andrew; Meeks, Kitty

DOI:
10.1016/j.dam.2018.02.008

Document Version
Peer reviewed version

Citation for published version (Harvard):
Treglown, A & Meeks, K 2018, 'On the complexity of finding and counting solution-free sets of integers', *Discrete Applied Mathematics*, vol. 243, pp. 219-238. https://doi.org/10.1016/j.dam.2018.02.008

ON THE COMPLEXITY OF FINDING AND COUNTING SOLUTION-FREE SETS OF INTEGERS

KITTY MEEKS AND ANDREW TREGLOWN

ABSTRACT. Given a linear equation $\mathcal{L}$, a set A of integers is $\mathcal{L}$-free if A does not contain any 'non-trivial' solutions to $\mathcal{L}$. This notion incorporates many central topics in combinatorial number theory such as sum-free and progression-free sets. In this paper we initiate the study of (parameterised) complexity questions involving $\mathcal{L}$-free sets of integers. The main questions we consider involve deciding whether a finite set of integers A has an $\mathcal{L}$-free subset of a given size, and counting all such $\mathcal{L}$-free subsets. We also raise a number of open problems.

1. INTRODUCTION

Sets of integers which do not contain any solutions to some linear equation have received a lot of attention in the field of combinatorial number theory. Two particularly well-studied examples are *sum-free sets* (sets avoiding solutions to the equation $x+y=z$) and *progression-free sets* (sets that do not contain any 3-term arithmetic progression x, y, z or equivalently avoid solutions to the equation $x+z=2y$). A lot of effort has gone into determining the size of the largest solution-free subset of $\{1,\dots,n\}$ and other sets of integers, and into computing (asymptotically) the number of (maximal) solution-free subsets of $\{1,\dots,n\}$.

In this paper we initiate the study of the computational complexity of problems involving solution-free subsets. We are primarily concerned with determining the size of the largest subset of an arbitrary set of integers A which avoids solutions to a specified linear equation $\mathcal{L}$; in particular, we focus on sum-free and progression-free sets, but many of our results also generalise to larger families of linear equations. For suitable equations $\mathcal{L}$, we demonstrate that the problem of deciding whether A contains a solution-free subset of size at least k is NP-complete (see Section 2); we further show that it is hard to approximate the size of the largest solution-free subset within a factor $(1+\epsilon)$ (see Section 3), or to determine for a constant $c<1$ whether A contains a solution-free subset of size at least $c|A|$ (see Section 6). On the other hand, in Section 5 we see that the decision problem is fixed-parameter tractable when parameterised by either the cardinality of the desired solution-free set, or by the number of elements of A we can exclude from such a set. We also consider the complexity, with respect to various parameterisations, of counting the number of solution-free sets of a specified size (see Section 7): while there is clearly no polynomial-time algorithm in general, the problem is fixed-parameter tractable when parameterised by the number of elements we can exclude from A; we show that there is unlikely to be a fixed-parameter algorithm to solve the counting problem exactly when the size of the solution-free sets is taken as the parameter, but we give an efficient approximation algorithm for this setting. Finally, in Section 8 we consider all of these questions in a variant of the problem, where we specify that a given solution-free subset $B \subset A$ must be included in any solution.

Many of our results are based on the fact that we can set up polynomial-time reductions in both directions between our problem and (different versions of) the well-known hitting set problem for

Date: March 6, 2018.

The first author is supported by a Personal Research Fellowship from the Royal Society of Edinburgh, funded by the Scottish Government, and the second author is supported by EPSRC grant EP/M016641/1.

hypergraphs. In particular, in Section 2.1 we provide a construction that has several applications throughout the paper. In Section 4 we also derive some new lower-bounds on the size of the largest solution-free subset of an arbitrary set of integers for certain equations $\mathcal{L}$, which may be of independent interest. Our approach here utilises a trick of Alon and Kleitman [3] which transfers the problem into the setting of solution-free sets in cyclic groups.

Our aim is to provide a thorough introduction to the study of (parameterised) complexity questions involving $\mathcal{L}$-free sets of integers. As such, some of the results presented have straightforward proofs, such as the parameterised complexity results discussed in Section 5, whilst other proofs are more involved. However, even the simplest of our results lead to natural open questions. In Section 9 we collect together a number of open problems which we hope will stimulate further interest in the topic.

In the remainder of this section, we give some background on solution-free sets in Section 1.1 and review the relevant notions from the study of computational complexity in Section 1.2. In Section 1.3 we outline the main results of the paper.

1.1. **Background on solution-free sets.** Consider a fixed linear equation $\mathcal{L}$ of the form

$$a_1x_1 + \cdots + a_\ell x_\ell = b \tag{1}$$

where $a_1, \ldots, a_\ell, b \in \mathbb{Z}$. We say that $\mathcal{L}$ is *homogeneous* if $b = 0$. If

$$\sum_{i\in[k]} a_i = b = 0$$

then we say that $\mathcal{L}$ is *translation-invariant*. (Here $[k]$ denotes the set $\{1, \ldots, k\}$.) Let $\mathcal{L}$ be translation-invariant. Then notice that $(x, \ldots, x)$ is a 'trivial' solution of (1) for any x. More generally, a solution $(x_1, \ldots, x_k)$ to $\mathcal{L}$ is said to be *trivial* if there exists a partition $P_1, \ldots, P_\ell$ of $[k]$ so that:

(i) $x_i = x_j$ for every i, j in the same partition class P_r;
(ii) For each $r \in [\ell]$, $\sum_{i\in P_r} a_i = 0$.

A set A of integers is *$\mathcal{L}$-free* if A does not contain any non-trivial solutions to $\mathcal{L}$. If the equation $\mathcal{L}$ is clear from the context, then we simply say A is *solution-free*.

1.1.1. *Sum-free sets.* A set S (of integers or elements of a group) is *sum-free* if there does not exist x, y, z in S such that $x + y = z$. The topic of sum-free sets has a rich history spanning a number of branches of mathematics. In 1916 Schur [42] proved that, given $r \in \mathbb{N}$, if n is sufficiently large, then any r-colouring of $[n] := \{1, \ldots, n\}$ yields a monochromatic triple x, y, z such that $x + y = z$. (Equivalently, $[n]$ cannot be partitioned into r sum-free sets.) This theorem was followed by other seminal related results such as van der Waerden's theorem [46], and ultimately led to the birth of arithmetic Ramsey theory.

Paul Erdős had a particular affinity towards sum-free sets. In 1965 he [20] proved one of the cornerstone results in the subject: every set of n non-zero integers A contains a sum-free subset of size at least $n/3$. Employing the probabilistic method, Alon and Kleitman [3] improved this bound to $(n + 1)/3$ and further, Kolountzakis [33] gave a polynomial time algorithm for constructing such a sum-free subset. Then, using a Fourier-analytical approach, Bourgain [11] further improved the bound to $(n + 2)/3$ in the case when A consists of positive integers. Erdős [20] also raised the question of determining upper bounds for this problem: recently Eberhard, Green and Manners [18] asymptotically resolved this important classical problem by proving that there is a set of positive integers A of size n such that A does not contain any sum-free subset of size greater than $n/3+o(n)$. This result raises the question of whether one can decide *efficiently* whether a set A of non-negative integers contains a sum-free subset of size at least $c|A|$ for some $c > 1/3$. As we shall see in Section 6, the answer is likely to be no.

In Section 7 we consider the complexity, with respect to various parameterisations, of counting the number of sum-free sets of a specified size. A number of important questions concerning (counting) sum-free sets were raised in two papers of Cameron and Erdős [13, 14]. In [13], Cameron and Erdős conjectured that there are $\Theta(2^{n/2})$ sum-free subsets of $[n]$. Here, the lower bound follows by observing that the largest sum-free subset of $[n]$ has size $\lceil n/2 \rceil$; this is attained by the set of odds in $[n]$ and by $\{\lfloor n/2 \rfloor + 1, \ldots, n\}$. Then, for example, by taking all subsets of $[n]$ containing only odd numbers one obtains at least $2^{n/2}$ sum-free subsets of $[n]$. After receiving much attention, the Cameron–Erdős conjecture was proven independently by Green [27] and Sapozhenko [40]. Given a set A of integers we say $S \subseteq A$ is a *maximal sum-free subset of* A if S is sum-free and it is not properly contained in another sum-free subset of A. Cameron and Erdős [14] raised the question of how many maximal sum-free subsets there are in $[n]$. Very recently, this question has been resolved via a combinatorial approach by Balogh, Liu, Sharifzadeh and Treglown [7, 8].

Sum-free sets have also received significant attention with respect to groups. One highlight in this direction is work of Diananda and Yap [15] and Green and Ruzsa [28] that determines the size of the largest sum-free subset for every finite abelian group. In each case the largest sum-free set has size linear in the size of the abelian group. Another striking result in the area follows from Gowers' work on quasirandom groups. Indeed, Gowers [31] proved that there are non-abelian groups for which the largest sum-free subset has sublinear size, thereby answering a question of Babai and Sós [5]. See the survey of Tao and Vu [45] for a discussion on further problems concerning sum-free sets in groups.

1.1.2. *Progression-free sets.* A set S (of integers or elements of a group) is *progression-free* if there does not exist distinct x, y, z in S such that $x + y = 2z$. The study of progression-free sets has focused on similar questions to those relating to sum-free sets.

Unlike in the case of sum-free sets, one cannot ensure that every finite set of non-zero integers contains a progression-free subset of linear size. Indeed, a classical result of Roth [36] implies that the largest progression-free subset of $[n]$ has size $o(n)$. This has led to much interest in determining good bounds on the size of such a subset of $[n]$. See [10, 19, 29, 39] for the state-of-the-art lower and upper bounds for this problem.

Roth's theorem has been generalised in various directions; most famously via Szemerédi's theorem [44] which ensures that, if n is sufficiently large, every subset of $[n]$ of linear size contains arithmetic progressions of arbitrary length. Analogues of Roth's theorem have also been considered for finite abelian groups; see, for example, [12, 24, 34].

As in the case of sum-free sets, it is also natural to ask for the number of progression-free subsets of $[n]$. More generally, Cameron and Erdős [13] raised the question of how many subsets of $[n]$ do not contain an arithmetic progression of length k. Significant progress on the problem has recently been made in [6, 9, 41].

We remark that there has also been much work on $\mathcal{L}$-free sets other than the cases of sum-free and progression-free sets; see for example [32, 37, 38].

1.2. **Computational complexity.** In this paper we are concerned with determining which problems related to solution-free sets of integers are computationally tractable. In the first instance, we seek to classify decision problems as either belonging to the class P (i.e. being solvable in polynomial time) or being NP-hard and so unlikely to admit polynomial-time algorithms. For further background on computational complexity, the classes P and NP, and polynomial-time reductions we refer the reader to [26].

When dealing with sets of positive integers as input, it should be noted that the amount of space required to represent the input depends both on the cardinality of the set and on the magnitude of the numbers in the set. Given any finite set $A \subseteq \mathbb{Z}$, we write $\max(A)$ and $\min(A)$ for the elements of A whose values are maximum and minimum respectively; we further define $\max^*(A) := \max\{|a| :$

$a \in A\}$ and $\min^*(A) := \min\{|a| : a \in A\}$. We write $\text{size}(A)$ for the number of bits required to represent A, and note that there exist positive constants c_1 and c_2 such that

$$c_1 \max\{|A| \log(\min^*(A)), \log(\max^*(A))\} \leq \text{size}(A) \leq c_2 |A| \log(\max^*(A)).$$

We therefore consider a problem involving the set A to belong to P if it can be solved by an algorithm whose running time is bounded by a polynomial function of $\text{size}(A)$; note that this is true if and only if the running time is bounded by a polynomial function of $|A| \log(\max^*(A))$. If $A = \{a_1, \ldots, a_n\}$, we will assume that A is stored in such a way that, given i, we can read the element a_i in time $\mathcal{O}(\log(|a_i|))$.

There are two basic operations which we will need to consider in almost all of the algorithms and reductions discussed in this paper. First of all, we often need to determine whether a given ℓ-tuple $(x_1, \ldots, x_\ell)$ is a solution to the ℓ-variable linear equation $\mathcal{L}$. Note that we assume throughout that ℓ and all coefficients in $\mathcal{L}$ are constants, but $x_1, \ldots, x_\ell$ are taken to be part of the input; thus, allowing for the time required to carry out the necessary arithmetic operations, we can certainly determine whether $(x_1, \ldots, x_\ell)$ is a solution in time $\mathcal{O}(\log(\max_{1 \leq i \leq \ell} |x_i|))$.

Secondly, we will in many cases need to determine whether a given set $A \subseteq \mathbb{Z}$ is $\mathcal{L}$-free. We can do this in a naive way by considering all possible ℓ-tuples, and checking for each one whether it is a solution. Since there are $|A|^\ell$ possible ℓ-tuples, we see by the reasoning above that we can complete this procedure in time $\mathcal{O}\left(|A|^\ell \cdot \log^2(\max^*(A))\right)$. Note that, as ℓ is a constant, this is a polynomial-time algorithm in terms of $\text{size}(A)$.

In this paper, we will also discuss the parameterised complexity of various decision problems involving solution-free sets. Parameterised complexity provides a multivariate framework for the analysis of hard problems: if a problem is known to be NP-hard, so that we expect the running-time of any algorithm to depend exponentially on some aspect of the input, we can seek to restrict this exponential blow-up to one or more *parameters* of the problem rather than the total input size. This has the potential to provide an efficient solution to the problem if the parameter(s) in question are much smaller than the total input size. A parameterised problem with total input size n and parameter k is considered to be tractable if it can be solved by a so-called *fpt-algorithm*, an algorithm whose running time is bounded by $f(k) \cdot n^{\mathcal{O}(1)}$, where f can be any computable function. Such problems are said to belong to the complexity class FPT. The primary method for showing that a problem is unlikely to belong to FPT is to show that it is hard for some class $W[t]$ (where $t \geq 1$) in the W-heirarchy (see [23] for a formal definition of these classes). When reducing one parameterised problem to another in order to demonstrate the hardness of a parameterised problem, we have to be a little more careful than with standard NP-hardness reductions: as well as making sure that we can construct the new problem instance efficiently, we also need to ensure that the parameter value in the new problem depends only on the parameter value in the original problem.

Let Σ be a finite alphabet. We define a *parameterised decision problem* to be a pair (Π, κ) where $\Pi : \Sigma^* \to \{\text{YES}, \text{NO}\}$ is a function and $\kappa : \Sigma^* \to \mathbb{N}$ is a parameterisation (a polynomial-time computable mapping). Then, an *fpt-reduction from* (Π, κ) *to* (Π', κ') is an algorithm A such that

(1) A is an fpt-algorithm;
(2) given a yes-instance of Π as input, A outputs a yes-instance of Π', and given a no-instance of Π as input, A outputs a no-instance of Π';
(3) there is a computable function g such that, if I is the input to A, then $\kappa'(A(I)) \leq g(\kappa(I))$.

For further background on the theory of parameterised complexity we refer the reader to [17, 23]. We will also consider parameterised counting problems in Section 7, and relevant notions will be introduced at the start of this section.

1.3. **Summary of results.** We conclude the introduction by informally stating the main results of the paper, just in the cases of sum-free and progression-free sets:

- We prove that the problem of determining whether a finite set of integers contains a sum-free (or progression-free) subset of a given size is NP-complete (see Theorem 4).
- We prove that the problem of determining whether a finite set of integers A contains a sum-free subset covering a cth proportion of A is NP-complete for any fixed $c > 1/3$ (see Theorem 18). This complements the aforementioned result of Erdős [20] that ensures any such set A contains a sum-free subset of size at least $|A|/3$.
- In contrast, the problem of determining whether a finite set of integers A contains a progression-free subset covering a cth proportion of A is NP-complete for any fixed $c > 0$ (see Theorem 19).
- In Theorem 8 we show that it is unlikely one can efficiently approximate the size of the largest sum-free (and progression-free) subset in a finite set of integers.
- It is essentially trivial to see that the problem of determining whether a finite set of integers A contains a sum-free (or progression-free) subset of a given size k is in FPT, when parameterised by k (see Proposition 12). Using [23, Theorem 1.14], it is straightforward to conclude the same assertion holds if instead we parameterise by $|A| - k$.
- A result of Thurley [43] easily implies that the analogous problem, for *counting* the number of sum-free (or progression-free) subsets of A of a given size k, belongs to FPT when parameterised by $|A| - k$ (see Theorem 20) .
- Perhaps surprisingly, it is unlikely that this problem is in FPT when instead we parameterise by k (see Theorem 23).

2. The decision problem

In this section, we consider the following problem, where $\mathcal{L}$ is any fixed linear equation.

> $\mathcal{L}$-Free Subset
> *Input:* A finite set $A \subseteq \mathbb{Z}$ and $k \in \mathbb{N}$.
> *Question:* Does there exist an $\mathcal{L}$-free subset $A' \subseteq A$ such that $|A'| = k$?

We show that this problem is closely related to the well-known Hitting Set problem, and exploit this relationship to show that the problem is NP-complete for a large family of equations $\mathcal{L}$, including those defining both sum-free and progression-free sets. On the other hand, we see that the problem is polynomially solvable whenever $\mathcal{L}$ is a linear equation with only two variables.

We begin in Section 2.1 by showing how to construct an instance of $\mathcal{L}$-Free Subset that corresponds in a specific way to a given hypergraph; we will make use of this same construction to prove many results in later sections of the paper. We exploit this construction to give an NP-completeness proof in Section 2.2, before considering the two-variable case in 2.3.

2.1. **A useful construction.** In this section we describe the main construction we will exploit throughout the paper, and prove its key properties. Recall that an ℓ-uniform hypergraph is a hypergraph in which every edge has size exactly ℓ. Given any set X and $\ell \in \mathbb{N}$, we write X^ℓ for the set of ordered ℓ-tuples whose elements belong to X.

Lemma 1. *Let $\mathcal{L}$ be a linear equation $a_1x_1 + \cdots + a_\ell x_\ell = by$ where each $a_i \in \mathbb{N}$ and $b \in \mathbb{N}$ are fixed, and let $H = (V, E)$ be an ℓ-uniform hypergraph. Then we can construct in polynomial time a set $A \subseteq \mathbb{N}$ with the following properties:*

(1) A is the disjoint union of two sets A' and A'', where $|A'| = |V|$ and $|A''| = |E|$;
(2) there exist bijections $\phi_V : A' \to V$ and $\phi_E : A'' \to E$;

(3) for every $(x_1, \ldots, x_\ell, y) \in A^{\ell+1}$, we have that $(x_1, \ldots, x_\ell, y)$ is a non-trivial solution to $\mathcal{L}$ if and only if $x_1, \ldots, x_\ell \in A'$, $y \in A''$ and $\{\phi_V(x_1), \ldots, \phi_V(x_\ell)\} = \phi_E(y)$;
(4) $\log(\max(A)) = \mathcal{O}(|V|)$.

Proof. Write $a := \max_{1 \leq j \leq \ell} a_j$ and set $d := 2\ell a^2 b^2$. Let $V =: \{v_1, \ldots, v_n\}$ denote the vertex set of H. Define

$$A' := \{bd^i : \ i \in [n]\}$$

and

$$A'' := \left\{(a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}) : i_1 < i_2 < \cdots < i_\ell \text{ and } v_{i_1} v_{i_2} \ldots v_{i_\ell} \in E\right\}.$$

Further, define $\phi_V : A' \to V$ by setting $\phi_V(bd^i) = v_i$ for all $i \in [n]$, and note that ϕ_V is a well-defined bijection. We define $\phi_E : A'' \to E$ by setting $\phi_E(a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}) = v_{i_1} v_{i_2} \ldots v_{i_\ell}$ where $i_1 < i_2 < \cdots < i_\ell$ and $v_{i_1} v_{i_2} \ldots v_{i_\ell} \in E$; to see that ϕ_E is also a well-defined bijection it suffices to observe that, by the uniqueness of base-d representation of natural numbers, there is a unique way to write any $y \in A''$ in the form $a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}$. It follows from the bijectivity of ϕ_V and ϕ_E that we have defined A' so that for each vertex $v_i \in V$ there is a unique number $bd^i \in A'$ associated with it, and defined A'' so that for each edge $v_{i_1} v_{i_2} \ldots v_{i_\ell} \in E$ there is a unique number $(a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}) \in A''$ associated with it. Define $A := A' \cup A''$. Given H we produce A in time $\mathcal{O}((|V| + |E|) \log bd^{|V|}) = \mathcal{O}(n^{\ell+1})$.

Notice that conditions (1), (2) and (4) of the lemma immediately hold. To prove (3) note that it suffices to prove the following claim.

Claim. *The only non-trivial solutions $(x_1, x_2, \ldots, x_\ell, y)$ to $\mathcal{L}$ in A are such that each $x_j = bd^{i_j} \in A'$ for some $i_1 < i_2 < \cdots < i_\ell$ and $y = (a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}) \in A''$.*

To prove the claim it is helpful to consider the natural numbers working in base d. We will use the *coordinate notation* $[c_0, c_1, c_2, \ldots]$ to denote the natural number $c_0 d^0 + c_1 d^1 + c_2 d^2 + \ldots$. So with respect to this notation, each $bd^i \in A'$ has a zero in each coordinate except the ith coordinate, which takes value b. Each $(a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}) \in A''$ takes value a_j in its i_jth coordinate, and zero otherwise.

Suppose we have $x_1, \ldots, x_t \in A$ for some $t \in \mathbb{N}$. Define $\text{coord}(x_1, \ldots, x_t)$ to be the set of all integers $i \geq 0$ such that for at least one of the elements x_j in $\{x_1, \ldots, x_t\}$, the ith coordinate of x_j is non-zero. Note that (since d is sufficiently large compared with the a_i and b) we have $|\text{coord}(x)| = 1$ for all $x \in A'$ and $|\text{coord}(y)| = \ell$ for all $y \in A''$. Moreover, for any $x, x' \in A$, $\text{coord}(x) = \text{coord}(x')$ if and only if $x = x'$.

Suppose $(x_1, \ldots, x_\ell, y)$ is a non-trivial solution to $\mathcal{L}$ in A. (Note the choice of $\mathcal{L}$ ensures the only trivial solutions to $\mathcal{L}$ are such that $x_1 = \cdots = x_\ell = y$.) Crucially we defined d to be large with respect to the a_i and b. Thus, $\text{coord}(x_1, \ldots, x_\ell) = \text{coord}(a_1 x_1 + a_2 x_2 + \cdots + a_\ell x_\ell)$. That is, in coordinate notation, the coordinates of $(a_1 x_1 + a_2 x_2 + \cdots + a_\ell x_\ell)$ that are non-zero are precisely those coordinates that are non-zero in at least one of the x_j. So this gives us that $\text{coord}(x_1, \ldots, x_\ell) = \text{coord}(by)$.

If $y \in A'$ then $|\text{coord}(by)| = 1$. Note in this case we obtain a contradiction if $|\text{coord}(x_1, \ldots, x_\ell)| \geq 2$. So it must be the case that $x_1 = \cdots = x_\ell$ and $x_1 \in A'$. This means that there is some $i \in [n]$ such that $x_1 = \cdots = x_\ell = bd^i$; $y = bd^i$; and further $a_1 + \cdots + a_\ell = b$. Thus $(x_1, \ldots, x_\ell, y)$ is a trivial solution to $\mathcal{L}$, a contradiction to our assumption.

Therefore $y \in A''$. Write $y = a_1 d^{i_1} + a_2 d^{i_2} + \cdots + a_\ell d^{i_\ell}$ where $i_1 < i_2 < \cdots < i_\ell$. Suppose that all of the x_j lie in A''. Then since $\text{coord}(x_1, \ldots, x_\ell) = \text{coord}(by)$, we must have that $x_1 = \cdots = x_\ell = y$. This implies $a_1 + \cdots + a_\ell = b$ and hence $(x_1, \ldots, x_\ell, y)$ is a trivial solution to $\mathcal{L}$, a contradiction.

Next suppose there is at least one $x_j \in A''$ and at least one $x_{j'} \in A'$. Without loss of generality, we may assume that there is some $1 \leq \ell' \leq \ell - 1$ so that $x_1, \ldots, x_{\ell'} \in A''$ and $x_{\ell'+1}, \ldots, x_\ell \in A'$. Since $\text{coord}(x_1, \ldots, x_\ell) = \text{coord}(by)$, we must have that $x_1 = \cdots = x_{\ell'} = y$. Furthermore, as

$a_1x_1+\cdots+a_\ell x_\ell = by$ we have that $a_1x_1+\cdots+a_{\ell'}x_{\ell'} < by$ and thus $a_1+\cdots+a_{\ell'} < b$. In particular, $\text{coord}((b-a_1-\cdots-a_{\ell'})y) = \{i_1,\ldots,i_\ell\}$. Therefore, as $a_{\ell'+1}x_{\ell'+1}+\cdots+a_\ell x_\ell = (b-a_1-\cdots-a_{\ell'})y$, this implies $\text{coord}(a_{\ell'+1}x_{\ell'+1}+\cdots+a_\ell x_\ell) = \{i_1,\ldots,i_\ell\}$. However, since $\ell' \geq 1$ and $x_{\ell'+1},\ldots,x_\ell \in A'$, we have that $|\text{coord}(a_{\ell'+1}x_{\ell'+1}+\cdots+a_\ell x_\ell)| < \ell$, a contradiction.

Altogether this implies that each $x_j \in A'$. The claim, and therefore lemma, now immediately follows. □

A relationship between independent sets in H and $\mathcal{L}$-free subsets of A now follows easily.

Corollary 2. *Let $\mathcal{L}$ be a linear equation $a_1x_1+\cdots+a_\ell x_\ell = by$ where each $a_i \in \mathbb{N}$ and $b \in \mathbb{N}$ are fixed, and let $H = (V,E)$ be an ℓ-uniform hypergraph. Let A and A'' be as in Lemma 1 on input H and $\mathcal{L}$. Then, for any $k \in \mathbb{N}$, there is a one-to-one correspondence between independent sets of H of cardinality k and the $\mathcal{L}$-free subsets of A of cardinality $|A''|+k$ which contain all the elements of A''.*

Proof. The corollary follows immediately from Lemma 1. Indeed, let ϕ_V and ϕ_E be as in Lemma 1. Given an independent set I of H, note that $\phi_V^{-1}(I) \cup A''$ is an $\mathcal{L}$-free subset of A of size $|I|+|A''|$; by bijectivity of ϕ_V, the $\mathcal{L}$-free subsets corresponding to independent sets $I_1 \neq I_2$ are distinct. Further, given any $\mathcal{L}$-free subset $S \cup A''$ of A of size $|S|+|A''|$, we have that $\phi_V(S)$ is an independent set in H of size $|S|$; again, by bijectivity of ϕ_V, we obtain a unique independent set for each such $\mathcal{L}$-free subset. □

Finally, if we are only interested in the existence of independent sets, we can drop one of the conditions on the $\mathcal{L}$-free subsets.

Corollary 3. *Let $\mathcal{L}$ be a linear equation $a_1x_1+\cdots+a_\ell x_\ell = by$ where each $a_i \in \mathbb{N}$ and $b \in \mathbb{N}$ are fixed, and let $H = (V,E)$ be an ℓ-uniform hypergraph. Let A and A'' be as in Lemma 1 on input H and $\mathcal{L}$. Then, for any $k \in \mathbb{N}$, H contains an independent set of cardinality k if and only if A contains an $\mathcal{L}$-free subset of cardinality $|A''|+k$.*

Proof. By Corollary 2, it suffices to show that, if A contains an $\mathcal{L}$-free subset of cardinality $|A''|+k$, then in fact A contains such a subset which includes all elements of A''. To see this, let A_1 be an $\mathcal{L}$-free subset of A of size $|A''|+k$ which does not contain all elements of A''; we will show how to construct an $\mathcal{L}$-free subset of equal or greater size which does have this additional property.

Suppose $y \in A''$ such that $y \notin A_1$. By Lemma 1(3) there is a unique choice of $x_1,\ldots,x_\ell \in A$ such that $(x_1,\ldots,x_\ell,y)$ is a non-trivial solution to $\mathcal{L}$ in A. If one of these x_j does not lie in A_1 we add y to A_1 without creating a solution to $\mathcal{L}$. Otherwise, arbitrarily remove one of the x_j from A_1 and replace it with y. Repeating this process, we obtain an $\mathcal{L}$-free subset which contains A'' and is at least as large as A_1. □

2.2. **The case of three or more variables.** The next result shows that for a range of linear equations, $\mathcal{L}$-Free Subset is NP-complete. For example, the result includes the cases when $\mathcal{L}$ is $x+y=z$ (i.e. sum-free sets) and $x+y=2z$ (i.e. progression-free sets). To prove that $\mathcal{L}$-Free Subset is NP-hard we will use a reduction from the following NP-complete problem [26]. Recall that a *hitting set* S in a hypergraph H is a collection of vertices such that every edge in H contains at least one vertex from S.

ℓ-Hitting Set
Input: An ℓ-uniform hypergraph H and $s \in \mathbb{N}$.
Question: Does H contain a hitting set of size s?

Theorem 4. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + \cdots + a_\ell x_\ell = b$ where each $a_i \in \mathbb{N}$ and $b \in \mathbb{N}$ are fixed and $\ell \geq 2$. Then $\mathcal{L}$-Free Subset is* NP*-complete.*

Proof. Recall from the discussion in Section 1.2 that we can determine in time polynomial in size(A) whether a set A is $\mathcal{L}$-free, so $\mathcal{L}$-Free Subset is in NP. To show that the problem is NP-complete, we give a reduction from ℓ-Hitting Set.

Let (H, s) be an instance of ℓ-Hitting Set. We construct A and $A'' \subseteq A$ as in Lemma 1 under input H and $\mathcal{L}$ (taking time polynomial in size(A)). It suffices to show that H has a hitting set of size s if and only if A contains an $\mathcal{L}$-free subset of size $k := |A| - s$.

Observe that H has a hitting set of size s if and only if it has an independent set of size $|H| - s$; by Corollary 3, this holds if and only if A has an $\mathcal{L}$-free subset of size $|A''| + |H| - s = |A| - s = k$. □

Note that since Lemma 1 outputs a set A of *natural numbers*, we have actually proved the following stronger result.

Theorem 5. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + \cdots + a_\ell x_\ell = b$ where each $a_i \in \mathbb{N}$ and $b \in \mathbb{N}$ are fixed and $\ell \geq 2$. Then $\mathcal{L}$-Free Subset is* NP*-complete, even if the input set A is a subset of $\mathbb{N}$.*

2.3. **The two variable case.** For *any* linear equation $\mathcal{L}$ in two variables, it is straightforward to see that $\mathcal{L}$-Free Subset is in P. Our strategy here is to reduce *to* the problem of finding an independent set (rather than reducing from this problem as in the previous section) and to note that the graph we create must have a very specific structure.

Theorem 6. *Fix any linear equation $\mathcal{L}$ in two variables. Then $\mathcal{L}$-Free Subset is in* P*.*

Proof. Let $A = \{a_1, \ldots, a_n\} \subseteq \mathbb{Z}$, and let $k \in \mathbb{N}$. We now construct G to be the graph with vertex set A, where $a_ia_j \in E(G)$ precisely when (a_i, a_j) is a non-trivial solution to $\mathcal{L}$. Note that we can construct G in time bounded by a polynomial function of size(A). The construction of G ensures that a set $A' \subseteq A$ is $\mathcal{L}$-free if and only if A' is an independent set in G.

Notice that a vertex x could lie in a loop. However, in this case x is not adjacent to any other vertex in G. All other vertices in G have degree at most 2. Thus, G is a collection of vertex-disjoint paths, cycles, isolated vertices and loops. The largest independent set in both a path and an even cycle on t vertices is $\lceil t/2 \rceil$; the largest independent set in an odd cycle on t vertices is $\lfloor t/2 \rfloor$. So in time $\mathcal{O}(n)$ we can determine the size of the largest independent set in G, and thus whether A contains an $\mathcal{L}$-free subset of size k. □

3. Approximating the size of the largest $\mathcal{L}$-Free Subset

Thus far we have focused on decision problems involving solution-free sets ("Does the set A contain a solution-free set of a certain size?"), but it is also natural to consider a maximisation problem: "What is the size of the largest solution-free subset of A?" An efficient algorithm to answer the decision problem can clearly be used to solve the maximisation problem, as we can repeatedly run our decision algorithm with different target sizes; however, as we have demonstrated in Section 2 that in many cases such an algorithm is unlikely to exist, it makes sense to ask whether we can efficiently approximate the optimisation problem. We define the maximisation version of $\mathcal{L}$-Free Subset formally as follows.

MAXIMUM $\mathcal{L}$-FREE SUBSET
Input: A finite set $A \subseteq \mathbb{Z}$.
Question: What is the cardinality of the largest $\mathcal{L}$-free subset $A' \subseteq A$?

Given any instance I of an optimisation problem, we denote by $\text{opt}(I)$ the value of the optimal solution to I (so, for example, the cardinality of the largest solution-free subset). Given a constant $\rho > 1$, we say that an approximation algorithm for the maximisation problem has *performance ratio* ρ if, given any instance I of the problem, the algorithm will return a value x such that

$$1 \leq \frac{\text{opt}(I)}{x} \leq \rho.$$

Note that there is a trivial approximation algorithm for MAXIMUM SUM-FREE SUBSET with performance ratio 3: if we always return $|A|/3$ then, as $|A|/3 \leq \text{opt}(A) \leq |A|$, we have $1 \leq \frac{\text{opt}(A)}{|A|/3} \leq 3$ as required.

We might hope to improve on this to obtain, given arbitrary positive ϵ, an approximation algorithm for MAXIMUM $\mathcal{L}$-FREE SUBSET with performance ratio $1 + \epsilon$. However, we will show in this section that in certain cases this is no easier than solving the problem exactly. Specifically, we show that for a large family of 3-variable linear equations (including those defining sum-free and progression-free sets), there is no polynomial-time approximation scheme unless $\mathsf{P} = \mathsf{NP}$.

A *polynomial-time approximation scheme (PTAS)* for a maximisation problem is an algorithm which, given any instance I of the problem and a constant $\epsilon > 0$, returns, in polynomial-time, a value x such that

$$1 \leq \frac{\text{opt}(I)}{x} \leq 1 + \epsilon.$$

Note that the exponent of the polynomial is allowed to depend on ϵ.

The complexity class APX contains all optimisation problems (whose decision version belongs to NP) which can be approximated within some constant factor in polynomial time; this class includes problems which do not admit a PTAS unless $\mathsf{P} = \mathsf{NP}$, so one way to demonstrate that an optimisation problem is unlikely to admit a PTAS is to show that it is hard for the class APX. In order to show that a problem is APX-hard (and so does not admit a PTAS unless $\mathsf{P} = \mathsf{NP}$), it suffices to give a PTAS reduction from another APX-hard problem.

Definition. *Let Π_1 and Π_2 be maximisation problems. A* PTAS reduction from Π_1 to Π_2 *consists of three polynomial-time computable functions f, g and α such that:*

(1) for any instance I_1 of Π_1 and any constant error parameter ϵ, f produces an instance $I_2 = f(I_1, \epsilon)$ of Π_2;

(2) if $\epsilon > 0$ is any constant and y is any solution to I_2 such that $\frac{opt(I_2)}{|y|} \leq \alpha(\epsilon)$, then $x = g(I_1, y, \epsilon)$ is a solution to I_1 such that $\frac{opt(I_1)}{|x|} \leq 1 + \epsilon$.

We cannot immediately deduce results about the inapproximability of MAXIMUM $\mathcal{L}$-FREE SUBSET from Corollary 3 together the inapproximability of INDEPENDENT SET, as the cardinality of the largest $\mathcal{L}$-free subset in A will in general be dominated by the cardinality of A''. However, we can instead reduce from MAX IS-3, the problem of finding the size of a maximum independent set in a graph of maximum degree 3, which was shown to be APX-hard by Alimonti and Kann [2]. For this reduction we imitate the approach of Froese, Janj, Nichterlein and Niedermeier [25], who obtained a result analogous to Lemma 1 when reducing 3-HITTING SET to the problem of finding a maximum subset of points in general position.

Corollary 3 implies that we have a polynomial-time reduction from 3-IS to $\mathcal{L}$-FREE SUBSET (for suitable $\mathcal{L}$) in which $|A''| \leq \frac{3|V|}{2}$. Since it is also well-known that in any graph G on n vertices

with maximum degree at most Δ, every maximal independent set has cardinality at least $\frac{n}{\Delta+1}$ (if the independent set is smaller than this, there must be some vertex which does not have any neighbour in the set and so can be added to the independent set), it follows that for every maximal independent set U in G we have $|U| \geq \frac{|V|}{4} \geq \frac{1}{6}|A''|$.

Using this observation, we can now define a PTAS reduction from Max IS-3 to Maximum $\mathcal{L}$-Free Subset for certain 3-variable equations.

Lemma 7. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + a_2x_2 = by$, where $a_1, a_2, b \in \mathbb{N}$ are fixed. Then there is a PTAS reduction from* Max IS-3 *to* Maximum $\mathcal{L}$-Free Subset.

Proof. We define the functions f, g and α as follows.

First, we let f be the function which, given an instance G of Max IS-3 (where $G = (V, E)$) and any $\epsilon > 0$, outputs the set $A = \phi_V^{-1}(V) \cup \phi_E^{-1}(E) \subseteq \mathbb{N}$ described in Lemma 1; we know from Lemma 1 that we can construct this set in polynomial time.

Next suppose that B is an $\mathcal{L}$-free subset in A. We can construct in polynomial time a set $\widetilde{B}$, with $|\widetilde{B}| \geq |B|$, such that

(1) $\phi_E^{-1}(E) \subseteq \widetilde{B}$; and
(2) $\widetilde{B}$ is a maximal $\mathcal{L}$-free subset of A.

If B fails to satisfy the first condition, we can use the method of Corollary 3 to obtain a set with this property, and if the resulting set is not maximal we can add elements greedily until this condition is met. We now define g to be the function which, given an $\mathcal{L}$-free set $B \subseteq A$ and any $\epsilon > 0$, outputs $\phi_V(\widetilde{B} \setminus \phi_E^{-1}(E))$.

Finally, we define α to be the function $\epsilon \mapsto 1 + \frac{\epsilon}{7}$. Let us denote by $\mathrm{opt}(G)$ the cardinality of the maximum independent set in G, and by $\mathrm{opt}(A)$ the cardinality of the largest $\mathcal{L}$-free subset in A. Note that $\mathrm{opt}(A) = \mathrm{opt}(G) + |E|$. To complete the proof, it suffices to demonstrate that, whenever B is an $\mathcal{L}$-free subset in A such that $\frac{\mathrm{opt}(A)}{|B|} \leq \alpha(\epsilon) = 1 + \frac{\epsilon}{7}$, we have $\frac{\mathrm{opt}(G)}{|I|} \leq 1 + \epsilon$, where $I := \phi_V(\widetilde{B} \setminus \phi_E^{-1}(E))$. Observe that

$$\begin{aligned}
\frac{\mathrm{opt}(A)}{|B|} &\leq 1 + \frac{\epsilon}{7} \\
\Rightarrow \frac{\mathrm{opt}(A)}{|\widetilde{B}|} &\leq 1 + \frac{\epsilon}{7} \\
\Rightarrow \frac{|E| + \mathrm{opt}(G)}{|E| + |I|} &\leq 1 + \frac{\epsilon}{7} \\
\Rightarrow \frac{|E| + \mathrm{opt}(G)}{|I|} &\leq \left(1 + \frac{\epsilon}{7}\right)\left(\frac{|E| + |I|}{|I|}\right) \\
\Rightarrow \frac{\mathrm{opt}(G)}{|I|} &\leq \left(1 + \frac{\epsilon}{7}\right)\frac{|E|}{|I|} + \left(1 + \frac{\epsilon}{7}\right) - \frac{|E|}{|I|} \\
&= \frac{\epsilon}{7}\frac{|E|}{|I|} + 1 + \frac{\epsilon}{7}.
\end{aligned}$$

Since we know that $|E| \leq \frac{3|V|}{2}$ and, by our assumptions on maximality of $\widetilde{B}$ and hence I, we also know that $|I| \geq \frac{|V|}{4}$, it follows that $\frac{|E|}{|I|} \leq 6$. We can therefore conclude that

$$\frac{\mathrm{opt}(A)}{|B|} \leq 1 + \frac{\epsilon}{7} \Rightarrow \frac{\mathrm{opt}(G)}{|I|} \leq 6\frac{\epsilon}{7} + 1 + \frac{\epsilon}{7} = 1 + \epsilon,$$

as required. □

We now obtain our main inapproximability result as an immediate corollary.

Theorem 8. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + a_2x_2 = by$, where $a_1, a_2, b \in \mathbb{N}$ are fixed. Then* Maximum $\mathcal{L}$-Free Subset *is* APX*-hard.*

4. $\mathcal{L}$-free subsets of arbitrary sets of integers

In much of the rest of this paper, we prove complexity results which hold whenever we can guarantee that our input set A will contain a reasonably large $\mathcal{L}$-free subset. We already know that this is the case for sum-free subsets (in which case an arbitrary input set A of non-zero integers must contain a sum-free subset of size at least $(|A|+1)/3$); in this section we extend this result to a much larger family of linear equations, proving that, given any homogeneous non-translation-invariant linear equation $\mathcal{L}$, every finite set of non-zero integers contains an $\mathcal{L}$-free set of linear size. Note that a homogeneous linear equation $\mathcal{L}$ is non-translation-invariant if and only if it can be written in the form $a_1x_1 + \cdots + a_kx_k = b_1y_1 + \cdots + b_\ell y_\ell$ for some fixed $a_i, b_i \in \mathbb{N}$ where $a_1 + \cdots + a_k \neq b_1 + \cdots + b_\ell$.

For this we will use a trick of Alon and Kleitman [3] which transfers the problem into the setting of solution-free sets in cyclic groups. We also utilise the following simple observation.

Observation 9. *Consider a homogeneous linear equation $\mathcal{L}$ of the form $a_1x_1 + \cdots + a_kx_k = b_1y_1 + \cdots + b_\ell y_\ell$ for some fixed $a_i, b_i \in \mathbb{N}$ where $a_1 + \cdots + a_k > b_1 + \cdots + b_\ell$. Then the interval*

$$I := \left[\left\lfloor \frac{(b_1 + \cdots + b_\ell)n}{a_1 + \cdots + a_k} \right\rfloor + 1, n\right]$$

is $\mathcal{L}$-free.

Note that Observation 9 is immediate since $(a_1 + \cdots + a_k)\min(I) > (b_1 + \cdots + b_\ell)\max(I)$.

Theorem 10. *Consider a non-translation-invariant homogeneous linear equation $\mathcal{L}$. There exists some $\lambda = \lambda(\mathcal{L}) > 0$ such that, if $n \in \mathbb{N}$ is sufficiently large, then any set $Z \subseteq \mathbb{Z} \setminus \{0\}$ so that $|Z| = n$ contains an $\mathcal{L}$-free subset of size more than λn.*

Proof. Suppose that $\mathcal{L}$ is of the form $a_1x_1 + \cdots + a_kx_k = b_1y_1 + \cdots + b_\ell y_\ell$ for some fixed $a_i, b_i \in \mathbb{N}$ where $a_1 + \cdots + a_k \neq b_1 + \cdots + b_\ell$. Observation 9 implies that there is some $\lambda' = \lambda'(\mathcal{L}) > 0$ such that, if $m' \in \mathbb{N}$ is sufficiently large, then $[m']$ contains an $\mathcal{L}$-free subset of size at least $\lambda' m'$. We say a subset S of a group G is *$\mathcal{L}$-free* if S contains *no* solutions to $\mathcal{L}$. Set $c := \max\{(a_1 + \cdots + a_k), (b_1 + \cdots + b_\ell)\}$.

Claim. *There is some $\lambda := \lambda'/(2c) > 0$ such that, if $m \in \mathbb{N}$ is sufficiently large, then $\mathbb{Z}_m$ contains an $\mathcal{L}$-free subset of size at least λm that does not contain the zero element.*

To prove the claim, suppose that $m \in \mathbb{N}$ is sufficiently large and define $m' := \lfloor m/c \rfloor$. Note a set $S \subseteq \{1, \ldots, m'\}$ is an $\mathcal{L}$-free subset of $\mathbb{Z}_m$ if and only if S is an $\mathcal{L}$-free subset of $[m']$. Indeed, suppose for a contradiction there is a solution $(x_1, \ldots, x_k, y_1, \ldots, y_\ell)$ to $\mathcal{L}$ in $\{1, \ldots, m'\} \subseteq \mathbb{Z}_m$ that is not a solution to $\mathcal{L}$ when viewed as a subset of $[m']$. So viewing $x_1, \ldots, x_k, y_1, \ldots, y_\ell$ as integers we have that $a_1x_1 + \cdots + a_kx_k \neq b_1y_1 + \cdots + b_\ell y_\ell$ however, $a_1x_1 + \cdots + a_kx_k \equiv b_1y_1 + \cdots + b_\ell y_\ell \mod m$. Thus the difference between $a_1x_1 + \cdots + a_kx_k$ and $b_1y_1 + \cdots + b_\ell y_\ell$ is at least m. This yields a contradiction since, by the definition of m', neither of these numbers is bigger than m.

Thus, as $[m']$ contains an $\mathcal{L}$-free subset of size at least $\lambda' m' \geq \lambda m$, $\mathbb{Z}_m$ contains an $\mathcal{L}$-free subset of size at least λm avoiding the zero element, proving the claim.

The rest of the proof modifies the argument presented in [4, Theorem 1.4.1] that shows every set of n non-zero integers contains a sum-free subset of size more than $n/3$. Let $n \in \mathbb{N}$ be sufficiently large and consider any set $Z = \{z_1, \ldots, z_n\}$ of non-zero integers. Let p be a prime so that $p > 2\max(Z)$. Since n is sufficiently large and $p > n$, by the claim we have that $\mathbb{Z}_p$ contains an $\mathcal{L}$-free subset S so that $|S| \geq \lambda p$ and additionally $0 \notin S$.

Choose an integer x uniformly at random from $\{1, 2, \ldots, p-1\}$, and define $d_1, \ldots, d_n$ by $d_i \equiv xz_i \mod p$ where $0 \leq d_i < p$. For every fixed $1 \leq i \leq n$, as x ranges over all numbers $1, 2, \ldots, p-1$, then d_i ranges over all non-zero elements of $\mathbb{Z}_p$. Therefore, $\mathbb{P}(d_i \in S) = |S|/(p-1) > \lambda$. So the expected number of elements z_i such that $d_i \in S$ is more than λn. Thus, there is some choice of x with $1 \leq x < p$ and a subset $Z' \subseteq Z$ of size $|Z'| > \lambda n$ such that $xz_i \pmod p \in S$ for all $z_i \in Z'$. Since S is $\mathcal{L}$-free in $\mathbb{Z}_p$, and $\mathcal{L}$ is homogeneous, this implies Z' is an $\mathcal{L}$-free set of integers, as desired. □

In the case when $\mathcal{L}$ is translation-invariant, Ruzsa [37] observed that the largest $\mathcal{L}$-free subset of $[n]$ has size $o(n)$. So one cannot prove an analogue of Theorem 10 for such equations $\mathcal{L}$.

The next result follows immediately from Theorem 10.

Theorem 11. *Consider a non-translation-invariant homogeneous linear equation $\mathcal{L}$. There exists some $\lambda = \lambda(\mathcal{L}) > 0$ such that every finite set $Z \subseteq \mathbb{Z} \setminus \{0\}$ contains an $\mathcal{L}$-free subset of size more than $\lambda|Z|$.*

Note that is necessary to restrict our attention to sets $Z \subseteq \mathbb{Z} \setminus \{0\}$ here since $Z := \{0\}$ does not contain a non-empty $\mathcal{L}$-free subset.

It is natural to ask how large $\lambda(\mathcal{L})$ can be in the previous theorem. Let $\mathcal{C}(\mathcal{L})$ denote the set of all positive reals κ so that Theorem 11 holds with κ playing the role of λ, and define $\mathcal{C}'(\mathcal{L})$ analogously now with respect to Theorem 10. We claim that $\mathcal{C}(\mathcal{L}) = \mathcal{C}'(\mathcal{L})$. It is immediate that $\mathcal{C}(\mathcal{L}) \subseteq \mathcal{C}'(\mathcal{L})$. To see that there is no $\lambda \in \mathcal{C}'(\mathcal{L}) \setminus \mathcal{C}(\mathcal{L})$ consider the following observation: Suppose $Z \subseteq \mathbb{Z} \setminus \{0\}$ is such that it does not contain an $\mathcal{L}$-free subset of size more than $\lambda|Z|$ for some $\lambda > 0$. Set $z := |Z|$. Then for every $n \in \mathbb{N}$ there is a set $Z' \subseteq \mathbb{Z} \setminus \{0\}$ of size zn such that it does not contain an $\mathcal{L}$-free subset of size more than $\lambda|Z'|$. Indeed, writing cZ as shorthand for $\{cz : z \in Z\}$, one can choose Z' to be the union of $c_1 Z, \ldots, c_n Z$ where the c_is are positive integers chosen to ensure the sets $c_i Z$ are pairwise disjoint. (Notice we required that $0 \notin Z$ to ensure this.)

Define

$$\kappa(\mathcal{L}) := \sup(\mathcal{C}(\mathcal{L})).$$

Write $\mathcal{L}$ as $a_1x_1 + \cdots + a_kx_k = b_1y_1 + \cdots + b_\ell y_\ell$ for some fixed $a_i, b_i \in \mathbb{N}$ where $a_1 + \cdots + a_k > b_1 + \cdots + b_\ell$. We remark that it is easy to check in the statement of Theorem 10, and therefore Theorem 11, one can set

$$\lambda(\mathcal{L}) = \frac{1}{2c}\left(1 - \frac{b_1 + \cdots + b_\ell}{a_1 + \cdots + a_k}\right) \tag{2}$$

where here we define $c := \max\{(a_1 + \cdots + a_k), (b_1 + \cdots + b_\ell)\}$. That is, $\kappa(\mathcal{L}) \geq \frac{1}{2c}\left(1 - \frac{b_1+\cdots+b_\ell}{a_1+\cdots+a_k}\right)$.

In the case when $\mathcal{L}$ is $x + y = z$ we know that $\kappa(\mathcal{L}) = 1/3$ and this supremum is attained. Indeed, recall that every set of n non-zero integers has a sum-free subset of size at least $(n+1)/3$ [3] whilst there are sets of positive integers A of size n such that A does not contain any sum-free subset of size greater than $n/3 + o(n)$ [18]. It would be interesting to determine $\kappa(\mathcal{L})$ for other equations $\mathcal{L}$.

Problem. *Determine $\kappa(\mathcal{L})$ for non-translation-invariant homogeneous linear equations $\mathcal{L}$.*

5. Parameterised complexity of the decision problem

Once one knows that a problem is NP-hard, there is interest in identifying the specific properties of a problem instance which contribute to the hardness; as mentioned in Section 1.2, one of the most natural ways to do this is to analyse the problem from the perspective of parameterised complexity, with the goal of establishing which parameterisations do or do not allow the design of fpt-algorithms. In this section we make two straightforward observations about the complexity of

$\mathcal{L}$-Free Subset with respect to two natural parameterisations, namely the number of elements in the sum-free subset (k) and the number of elements *not* in this subset ($|A| - k$).

First, consider parameterisation by k. Note that whenever $\mathcal{L}$ satisfies the conditions of Theorem 11 we know that there exists an explicit constant $\lambda = \lambda(\mathcal{L}) > 0$ such that any finite set $A \subset \mathbb{Z} \setminus \{0\}$ contains an $\mathcal{L}$-free subset of size at least $\lambda|A|$. Thus, if $|A|$ is large enough compared with k we can immediately answer yes; otherwise $|A|$ is bounded by a function of k and so we can solve the problem by brute force in time bounded by a function of k. This gives the following result.

Proposition 12. *Let $\mathcal{L}$ be a non-translation-invariant homogeneous linear equation. Then $\mathcal{L}$-Free Subset, parameterised by k, is in* FPT.

To deal with the dual parameterisation $|A| - k$, we use the fact that the following problem is known to belong to FPT [23, Theorem 1.14].

p-card-Hitting Set
Input: A hypergraph $G = (V, E)$ and $s \in \mathbb{N}$.
Parameter: $s + d$, where $d = \max_{e \in E} |e|$.
Question: Does G contain a hitting set of cardinality s?

The strategy is then to reduce $\mathcal{L}$-Free Subset to **p**-card-Hitting Set.

Lemma 13. *Let $\mathcal{L}$ be any fixed linear equation with ℓ variables, and let $A \subseteq \mathbb{Z}$ be finite. Then we can construct, in time polynomial in size(A), a hypergraph G on $|A|$ vertices in which every edge contains at most ℓ vertices, such that there is a one-to-one correspondence between $\mathcal{L}$-free subsets of A of cardinality k and hitting sets in G of cardinality $|A| - k$.*

Sketch proof. Suppose without loss of generality that $\mathcal{L}$ is of the form $a_1x_1 + \cdots + a_\ell x_\ell = b$, where $a_1, \ldots, a_\ell, b \in \mathbb{Z}$. Let G be the hypergraph with vertex set A and edge set

$$E := \{\{x_1, \ldots, x_\ell\} : \ (x_1, \ldots, x_\ell) \text{ is a non-trivial solution to } \mathcal{L}\}.$$

Note that $x_1, \ldots, x_\ell$ are not necessarily all distinct, so while every edge in E contains *at most* ℓ vertices, an edge may contain strictly fewer than ℓ vertices. There is then a one-to-one correspondence between hitting sets of cardinality k in G and $\mathcal{L}$-free subsets of cardinality $|A| - k$ in A. □

The fixed parameter tractability of our problem with respect to the parameter $|A| - k$ now follows immediately.

Theorem 14. *Let $\mathcal{L}$ be any fixed linear equation. Then $\mathcal{L}$-Free Subset, parameterised by $|A| - k$, belongs to* FPT.

Proposition 12 and Theorem 14 raise the question of what other natural parameterisations of $\mathcal{L}$-Free Subset yield tractability. We discuss this further in Section 9.

6. $\mathcal{L}$-free subsets covering a given fraction of elements

We know, by Theorem 4, that there is unlikely to be a polynomial time algorithm to decide whether a set A has an $\mathcal{L}$-free subset of size k, for arbitrary $k \in \mathbb{N}$. It is therefore natural to ask whether we can efficiently solve a restricted version of the problem in which we want to determine whether a finite set A of (non-zero) integers contains an $\mathcal{L}$-free subset that houses some fixed proportion of the elements of A. Given any linear equation $\mathcal{L}$ and $0 < \varepsilon < 1$, we define the following problem.

ε-$\mathcal{L}$-Free Subset
Input: A finite set $A \subseteq \mathbb{Z} \setminus \{0\}$.
Question: Does there exist an $\mathcal{L}$-free subset $A' \subseteq A$ such that $|A'| \geq \varepsilon|A|$?

In the case when $\mathcal{L}$ is $x+y=z$ we refer to ε-$\mathcal{L}$-Free Subset as ε-Sum-Free Subset. Note that ε-$\mathcal{L}$-Free Subset concerns finite sets of non-zero integers A; thus, the definition of $\kappa(\mathcal{L})$ (given in Section 4) immediately implies that ε-$\mathcal{L}$-Free Subset is in P for all $\varepsilon \leq \kappa(\mathcal{L})$, as in this case every instance is a yes-instance.

Further, recall from Section 1.2 that, given any fixed linear equation $\mathcal{L}$, we can decide in time polynomial in size(A') whether a set $A' \subseteq \mathbb{Z}$ is $\mathcal{L}$-free, so ϵ-$\mathcal{L}$-Free Subset clearly belongs to NP.

We will show in Section 6.2 that, for certain choices of $\mathcal{L}$ and ϵ, the ϵ-$\mathcal{L}$-Free Subset problem is no easier than $\mathcal{L}$-Free Subset. For this, we will actually restrict our attention to the case when we have input set $A \subseteq \mathbb{N}$. In this case, we need to be able to add elements to A without creating any additional solutions; we prove results about this in Section 6.1.

6.1. **Extending sets without creating additional solutions.** In Section 6.2, and also later in Section 7, we will make use of the following lemma, which allows us to extend sets without creating additional solutions to an equation.

Lemma 15. *Suppose $\mathcal{L}$ is a linear equation $ax+by=cz$ where $a,b,c \in \mathbb{N}$ are fixed and $a+b \neq c$. Suppose $A \subseteq \mathbb{N}$ is a finite set and $t \in \mathbb{N}$ so that $t > |A|$. Then there is a set $B \subseteq \mathbb{N}$ such that:*

(i) $|B| = t$;
(ii) $A \subseteq B$;
(iii) *the only solutions to $\mathcal{L}$ in B lie in A;*
(iv) $\max(B) = \mathcal{O}\left(t\,(\max(A))^2\right)$.

Moreover, B can be computed in time polynomial in size(A) and t.

Proof. Write $m := \max(A)$ and set $\tau := \min\{\frac{c}{a+b}, \frac{a+b}{c}\}$. Define $N \in \mathbb{N}$ to be the smallest natural number so that $\lfloor \tau N \rfloor \geq 2(a+b+c)m$ and $N - \lfloor \tau N \rfloor \geq t$. This choice of N means that $\lfloor \tau(N-1) \rfloor < 2(a+b+c)m$ or $(N-1) - \lfloor \tau(N-1) \rfloor < t$. Thus,

$$N \leq \max\left\{\frac{t}{1-\tau}, \frac{2(a+b+c)m}{\tau}\right\} + 1. \tag{3}$$

By Observation 9 and the choice of N, $[N]$ contains an $\mathcal{L}$-free subset I' so that

$$I' \subseteq [\lfloor \tau N \rfloor + 1, N]$$

and $|I'| = t - |A|$.

Chebyshev's theorem implies that there is a prime p so that $abcm < p < 2abcm$. Set $I'' := pI'$ and let $B := A \cup I''$. Note that we can clearly determine N and hence construct I' in time bounded by a polynomial function of size(A) and t. We can determine an appropriate value for p (and then construct I'') by exhaustively searching the specified interval and testing for primality in polynomial time (using the AKS test [1]). This set immediately satisfies (ii) and, since $p > m$, A and I'' are disjoint so (i) is satisfied. Moreover, it follows from (3) and the choice of p that $\max(B) \leq \max\left\{\frac{2abctm}{1-\tau}, \frac{4m^2abc(a+b+c)}{\tau}\right\} + 2abcm = \mathcal{O}\left(t\,(\max(A))^2\right)$, so (iv) is satisfied.

To see that (iii) is satisfied, we first observe that there are no solutions to $\mathcal{L}$ in I''. Since $\min(I'') > 2(a+b+c)m$ it is easy to check that there are no solutions to $\mathcal{L}$ in B which consist of two elements from A and one element from I''. Suppose there is a solution to $\mathcal{L}$ in B which consists of two elements z_1, z_2 from I'' and one element z_3 from A. Consider the case when $az_1 + bz_2 = cz_3$

(the other cases follow identically). Since every element of I'' is divisible by p we have that p divides cz_3. So as $c < p$ this implies p must divide z_3. However, no element of A is divisible by p since $\max(A) = m < p$, a contradiction. Hence B satisfies condition (iii). This completes the proof. □

We can also prove an analogous result for equations $\mathcal{L}$ of the form $ax + by = cz$ where $a, b, c \in \mathbb{N}$ are fixed and $a + b = c$. To do so, we will need the following fact.

Fact 16. *Suppose $\mathcal{L}$ is a linear equation $ax + by = cz$ where $a, b, c \in \mathbb{N}$ are fixed; $a + b = c$; and $a \leq b$. Given any $x_1 < x_2 < x_3$ that form a solution (x, y, z) to $\mathcal{L}$ in $\mathbb{N}$, we have that x_2 plays the role of z and $cx_2 > ax_3$.*

Proof. Consider any $x_1 < x_2 < x_3$ that form a solution (x, y, z) to $\mathcal{L}$ in $\mathbb{N}$. Note that since $a+b = c$, we have $cx_3 > \max\{(ax_1 + bx_2), (ax_2 + bx_1)\}$. Thus x_3 cannot play the role of z. Further, x_2 must play the role of z. Indeed, otherwise x_1 plays the role of z and then we have $ax + by > cz$, a contradiction. Altogether this implies that $cx_2 > ax_3$. □

Lemma 17. *Suppose $\mathcal{L}$ is a linear equation $ax + by = cz$ where $a, b, c \in \mathbb{N}$ are fixed and $a + b = c$. Suppose $A \subseteq \mathbb{N}$ is a finite set and $t \in \mathbb{N}$. Then there is a set $B \subseteq \mathbb{N}$ such that:*

(i) $|B| = |A| + t$;
(ii) $A \subseteq B$;
(iii) *the only non-trivial solutions to $\mathcal{L}$ in B lie in A;*
(iv) $\max(B) = 2\,(\max(A))\,c^t$.

Moreover, B can be computed in time polynomial in t and size(A).

Proof. Without loss of generality assume that $a \leq b$. Note that (x, y, z) is a non-trivial solution to $\mathcal{L}$ if and only if (x, y, z) is a solution to $\mathcal{L}$ with x, y, z distinct.

Set $m := \max(A)$, and define $A' := \{c^i \cdot 2m \ : \ i \in [t]\}$; we can clearly construct B in time polynomial in t and size(A). We claim that $B := A \cup A'$ is our desired set. Certainly (i), (ii) and (iv) follow immediately.

We now prove (iii). Suppose $x_1 < x_2 < x_3$ form a solution (x, y, z) to $\mathcal{L}$ in A'. Then by Fact 16 we must have that $cx_2 > ax_3$. However, by definition of A', $ax_3 \geq x_3 \geq cx_2$, a contradiction. So A' does not contain any non-trivial solutions to $\mathcal{L}$. The same argument shows that there are no non-trivial solutions to $\mathcal{L}$ in B which contain two elements from A' and one element from A. Finally suppose $x_1 < x_2 < x_3$ form a solution (x, y, z) to $\mathcal{L}$ in B where $x_1, x_2 \in A$ and $x_3 \in A'$. As before we must have that $cx_2 > ax_3$. However, $ax_3 > acm \geq cx_2$ by definition of A', a contradiction. This proves (iii). □

Lemma 17 will be applied in the next subsection to prove that for any equation $\mathcal{L}$ as in its statement, ε-$\mathcal{L}$-Free Subset is NP-complete for any $0 < \varepsilon < 1$.

6.2. **Hardness of ε-$\mathcal{L}$-Free Subset.** In this section we show that, in two specific cases, ε-$\mathcal{L}$-Free Subset is NP-complete. We begin with the case of sum-free subsets. Note that if $\varepsilon \leq \frac{1}{3}$ then the problem is trivially in P as, by the aforementioned result of Erdős [20], the answer is always "yes"; also if $\varepsilon = 1$ then it suffices to check whether the input set is sum-free (which can be done in polynomial time). We now demonstrate that the problem is NP-complete for all other values of ε. Recall that, given a set $X \subseteq \mathbb{N}$ and $y \in \mathbb{N}$, we write yX as shorthand for $\{yx : x \in X\}$.

Theorem 18. *Given any rational $1/3 < \varepsilon < 1$, ε-Sum-Free Subset is NP-complete.*

Proof. Recall that ε-Sum-Free Subset belongs to NP. To show that the problem is NP-hard, we describe a reduction from Sum-Free Subset (restricted to inputs $A \subseteq \mathbb{N}$), shown to be NP-hard in Theorem 5.

Suppose that (A, k) is an instance of Sum-Free Subset where $A \subseteq \mathbb{N}$. We will define a set $B \subseteq \mathbb{N}$ such that B has a sum-free subset of size at least $\varepsilon|B|$ if and only if A has a sum-free subset of size k. The construction of B depends on the value k.

First suppose that $k \leq \varepsilon|A|$. Set

$$d := \left\lceil \frac{\varepsilon|A| - k}{1 - \varepsilon} \right\rceil,$$

so that d is the least positive integer such that $\varepsilon(|A| + d) \leq k + d$ and hence $\lceil \varepsilon(|A| + d) \rceil = k + d$. Note that $d = \mathcal{O}(|A|)$. By Lemma 15 we can construct, in time polynomial in size(A) and d, a set $B \subseteq \mathbb{N}$ of size $|A| + d$ such that B has a sum-free subset of size $k + d$ if and only if A has a sum-free subset of size k. By the choice of d, we know that B has a sum-free subset of size $k + d$ if and only if B has a sum-free subset of size at least $\varepsilon(|A| + d)$; so B is a yes-instance for ε-Sum-Free Subset if and only if (A, k) is a yes-instance for Sum-Free Subset.

Now suppose $k > \varepsilon|A|$. The result of Eberhard, Green and Manners [18] implies that there is a set $S \subseteq \mathbb{N}$ such that the largest sum-free subset of S has size precisely $\varepsilon'|S|$ where $1/3 < \varepsilon' < \varepsilon$. In particular, through an exhaustive search, one can construct such a set S. Crucially, S is independent of our input (A, k) (so size(S), $|S|$ and $\max(S)$ are all fixed constants).

Set

$$r := \left\lceil \frac{k - \varepsilon|A|}{(\varepsilon - \varepsilon')|S|} \right\rceil,$$

and note that $r = \mathcal{O}(|A|)$. Set $m := \max(A)$, $m' := \max(S)$, and define $d_i := 3^i m (m')^{i-1}$ for each $1 \leq i \leq r$. Note that $\log d_r = \mathcal{O}(r + \log m)$, so for each $1 \leq i \leq r$, we can represent the set $d_i S$ in space $\mathcal{O}(\text{size}(A))$. Now define

$$A^* := A \cup \bigcup_{1 \leq i \leq r} d_i S.$$

The choice of the d_i ensures the only solutions to $x + y = z$ in A^* are such that $x, y, z \in A$ or $x, y, z \in d_i S$ for some $i \in [r]$. The largest sum-free set in $d_i S$ is of size $\varepsilon'|S| = \varepsilon'|d_i S|$. Define $k^* := k + r\varepsilon'|S|$, and observe that A has a sum-free subset of size k if and only if A^* has a sum-free subset of size k^*. By definition of r, $r\varepsilon'|S| \leq r\varepsilon|S| + \varepsilon|A| - k$, so we see that $k^* \leq \varepsilon(|A| + r|S|) = \varepsilon|A^*|$. Now we can argue precisely as in the first case: from A^* one can construct a set B in time polynomial in size(A) so that A^* has a sum-free subset of size k^* if and only if B has a sum-free subset of size at least $\varepsilon|B|$. In particular, B will be a yes-instance for ε-Sum-Free Subset if and only if (A, k) is a yes-instance for Sum-Free Subset, as required. □

We are also able to prove an NP-completeness result in the only other cases of three-variable equations $\mathcal{L}$ where $\kappa(\mathcal{L})$ is known, using a slight variation on the method of Theorem 18. In particular, the following result covers the case of progression-free sets. Recall that if $\mathcal{L}$ is translation-invariant, Ruzsa [37] observed that the largest $\mathcal{L}$-free subset of $[n]$ has size $o(n)$ (and so $\kappa(\mathcal{L}) = 0$).

Theorem 19. *Consider any rational $0 < \varepsilon < 1$ and let $\mathcal{L}$ denote the equation $ax + by = cz$ where $a, b, c \in \mathbb{N}$ and $a + b = c$. Then ε-$\mathcal{L}$-Free Subset is NP-complete.*

Proof. Fix ε and $\mathcal{L}$ as in the statement of the theorem; we will assume without loss of generality that $b \geq a$. Recall that ε-$\mathcal{L}$-Free Subset is in NP. To show NP-hardness, we once again give a reduction from $\mathcal{L}$-Free Subset (restricted to inputs $A \subseteq \mathbb{N}$), shown to be NP-hard in Theorem 5.

Suppose that (A, k) is an instance of $\mathcal{L}$-Free Subset, where $A \subseteq \mathbb{N}$. We will define a set $B \subseteq \mathbb{N}$ such that B has an $\mathcal{L}$-free subset of size at least $\varepsilon|B|$ if and only if A has an $\mathcal{L}$-free subset of size k. The construction of B depends on the value k.

First suppose that $k \leq \varepsilon|A|$. As in the proof of Theorem 18, we define d so that $\lceil \varepsilon(|A|+d) \rceil = k+d$. By Lemma 17 we can construct, in time bounded by a polynomial function of size(A), a set $B \subseteq \mathbb{N}$ of size $|A|+d$ such that (by conditions (i)–(iii) of the lemma) B has an $\mathcal{L}$-free subset of size $k+d$ if and only if A has an $\mathcal{L}$-free subset of size k. By our choice of d, this means that B is a yes-instance to ε-$\mathcal{L}$-Free Subset if and only if (A,k) is a yes-instance to $\mathcal{L}$-Free Subset.

Now suppose $k > \varepsilon|A|$. Since the largest $\mathcal{L}$-free subset of $[n]$ has size $o(n)$, we can find by exhaustive search a set $S \subseteq \mathbb{N}$ such that the largest $\mathcal{L}$-free subset of S has size $\varepsilon'|S|$ for some $0 < \varepsilon' < \varepsilon$. Crucially, S is independent of our input (A,k) (so $|S|$, $\max(S)$ and size(S) are all fixed constants).

As in the proof of Theorem 18, we set

$$r := \left\lceil \frac{k - \varepsilon|A|}{(\varepsilon - \varepsilon')|S|} \right\rceil.$$

Set $m := \max(A)$, $m' := \max(S)$, and define $d_i := 3^i c^i m (m')^{i-1}$ for each $1 \leq i \leq r$. Note that, for each d_i, we can represent the set $d_i S$ in space $\mathcal{O}(\text{size}(A))$. Observe also that $\max(A) < \min(d_1 S)$ and, for $1 \leq i \leq r-1$, $\max(d_i S) < \min(d_{i+1} S)$.

Now set

$$A^* := A \cup \bigcup_{1 \leq i \leq r} d_i S.$$

We claim that the only solutions to $ax + by = cz$ in A^* are such that $x, y, z \in A$ or $x, y, z \in d_i S$ for some $i \in [r]$. To see this, first suppose there are $x_1 < x_2 < x_3$ in A^* so that $x_3 \in d_i S$ for some $i \in [r]$, $x_2 \notin d_i S$, and x_1, x_2, x_3 form a solution to $\mathcal{L}$. Then by Fact 16 we have that $cx_2 > ax_3$. However, we also know that, if $i > 1$, $ax_3 \geq x_3 \geq d_i = 3cm' d_{i-1} > cm' d_{i-1} \geq cx_2$; if $i = 1$ then $ax_3 \geq 3acm > cm \geq cx_2$. In either case this gives a contradiction. Next suppose there are $x_1 < x_2 < x_3$ in A^* so that $x_2, x_3 \in d_i S$ for some $i \in [r]$, $x_1 \notin d_i S$, and x_1, x_2, x_3 form a solution to $\mathcal{L}$. Suppose $i > 1$. Then d_i divides x_2 and x_3 and so, by Fact 16, d_i divides ax_1 or bx_1. In particular, we have that $bx_1 \geq d_i$. However, since $x_1 \notin d_i S$, we have that $x_1 \leq d_{i-1} m'$ and so $bx_1 \leq cm' d_{i-1} < 3cm' d_{i-1} = d_i$, a contradiction. The case $i = 1$ yields an analogous contradiction. Altogether this indeed proves the only solutions (x, y, z) to $\mathcal{L}$ in A^* are such that $x, y, z \in A$ or $x, y, z \in d_i S$ for some $i \in [r]$.

Now we can continue as in the proof of Theorem 18. Note that the largest $\mathcal{L}$-free set in $d_i S$ is of size $\varepsilon'|S| = \varepsilon'|d_i S|$, and set $k^* := k + r\varepsilon'|S|$, so that A has an $\mathcal{L}$-free subset of size k if and only if A^* has an $\mathcal{L}$-free subset of size k^*. By definition of r, we have $k^* \leq \varepsilon(|A| + r|S|) = \varepsilon|A^*|$, so we can now argue precisely as in the first case to obtain a set B such that B is a yes-instance to ε-$\mathcal{L}$-Free Subset if and only if (A,k) is a yes-instance to $\mathcal{L}$-Free Subset. □

7. Counting solution-free subsets of a given size

As mentioned in the introduction, much of the previous work on sum-free (and more generally solution-free) sets has focused on counting problems. In this section we consider such questions in the (parameterised) complexity setting. Consider the following counting problem.

#Given-Size $\mathcal{L}$-Free Subset
Input: A finite set $A \subseteq \mathbb{Z}$ and $k \in \mathbb{N}$.
Question: How many $\mathcal{L}$-free subsets of A have cardinality exactly k?

It is clear that, whenever $\mathcal{L}$ satisfies the conditions of Theorem 4, there cannot be any polynomial-time algorithm for this problem unless $\mathsf{P} = \mathsf{NP}$, as such an algorithm would certainly tell us whether

or not the number of such subsets is zero and hence solve the decision problem. However, it is interesting to consider the complexity of the counting problem with respect to the parameterisations k and $|A| - k$, as we saw in Section 5 that the decision problem is tractable with respect to both of these parameterisations.

In Section 7.2 we show that the counting problem is in FPT when parameterised by $|A|-k$; similar to the proof of Theorem 14, this result relies on a reduction to a counting version of Hitting Set. In contrast, we show in Section 7.3 that, for certain equations $\mathcal{L}$, the counting problem is unlikely to admit an fpt-algorithm when parameterised by k; however, in many cases there is an efficient parameterised algorithm to solve this problem *approximately*, as we will see in Section 7.4. We begin in Section 7.1 with some background on the theory of parameterised counting complexity.

7.1. **Parameterised counting complexity.** We make use of the theory of parameterised counting complexity developed by Flum and Grohe [22, 23]. Let Σ be a finite alphabet. A parameterised counting problem is formally defined to be a pair (Π, κ) where $\Pi : \Sigma^* \to \mathbb{N}_0$ is a function and $\kappa : \Sigma^* \to \mathbb{N}$ is a parameterisation (a polynomial-time computable mapping). Flum and Grohe define two types of parameterised counting reductions, *fpt parsimonious reductions* and *fpt Turing reductions*. The latter is more flexible than the former, as it does not require us to preserve the number of witnesses as we tranform between problems; rather we must be able to compute the number of witnesses in one problem using information about the number of witnesses in one or more instances of the other problem, which allows us to make use of several standard techniques for counting reductions (such as polynomial interpolation and matrix inversion).

Definition. *An fpt Turing reduction from (Π, κ) to (Π', κ') is an algorithm A with an oracle to Π' such that*

(1) A computes Π,
(2) A is an fpt-algorithm with respect to κ, and
(3) there is a computable function $g : \mathbb{N} \to \mathbb{N}$ such that for all oracle queries "$\Pi'(I') = ?$" posed by A on input I we have $\kappa'(I') \leq g(\kappa(I))$.

In this case we write $(\Pi, \kappa) \leq_{\mathrm{T}}^{\mathrm{fpt}} (\Pi', \kappa')$.

There is an analogue of the W-hierarchy for counting problems; in order to demonstrate that a parameterised counting problem is unlikely to belong to FPT it suffices to show that it is hard (with respect to fpt-Turing reductions) for the first level of this heirarcy, #W[1] (see [23] for the formal definition of the class #W[1]).

A parameterised counting problem is considered to be efficiently approximable if it admits a *fixed parameter tractable randomised approximation scheme (FPTRAS)*, which is defined as follows:

Definition. *A fixed parameter tractable randomised approximation scheme (FPTRAS) for a parameterised counting problem (Π, κ) is a randomised approximation scheme that takes an instance I of Π (with $|I| = n$), and rational numbers $\varepsilon > 0$ and $0 < \delta < 1$, and in time $f(\kappa(I)) \cdot g(n, 1/\varepsilon, \log(1/\delta))$ (where f is any computable function, and g is a polynomial in n, $1/\varepsilon$ and $\log(1/\delta)$) outputs a rational number z such that*

$$\mathbb{P}[(1-\varepsilon)\Pi(I) \leq z \leq (1+\varepsilon)\Pi(I)] \geq 1 - \delta.$$

7.2. **Parameterisation by the number of elements not included in the solution-free set.** In this section we show that the counting problem, parameterised by $|A| - k$, is in FPT. This is a straightforward extension of the argument used for the decision problem in Section 5: since there is a one-to-one correspondence between $\mathcal{L}$-free subsets of cardinality k and hitting sets of cardinality $|A|-k$ in the construction described in Lemma 13, we can also make use of parameterised algorithms for *counting* hitting sets to count $\mathcal{L}$-free subsets. Thurley [43] describes an fpt-algorithm for the following counting version of the problem.

#**p**-card-Hitting Set
Input: A hypergraph $G = (V, E)$ and $s \in \mathbb{N}$.
Parameter: $s + d$, where $d = \max_{e \in E} |e|$.
Question: How many hitting sets in G have cardinality exactly s?

As for Theorem 14, our result now follows immediately.

Theorem 20. *Let $\mathcal{L}$ be any fixed linear equation. Then* #Given-Size $\mathcal{L}$-Free Subset*, parameterised by $|A| - k$, belongs to* FPT*.*

7.3. **Parameterisation by the cardinality of the solution-free set.** In contrast with the positive result in the previous section, we now show that there is unlikely to be an fpt-algorithm with respect to the parameter k to solve #Given-Size $\mathcal{L}$-Free Subset. To do this, we give an fpt-Turing reduction from the following problem, which can easily be shown to be #W[1]-hard by means of a reduction from **p-#**Clique (shown to be #W[1]-hard in [22]), along the same lines as the proof of the W[1]-hardness of **p-**Multicolour Clique in [21].

p-#Multicolour Clique
Input: A graph $G = (V, E)$, and a partition of V into k sets $V_1, \ldots, V_k$.
Parameter: k
Question: How many k-vertex cliques in G contain exactly one vertex from each set $V_1, \ldots, V_k$?

When reducing from **p-**Multicolour Clique or its counting version, it is standard practice to assume that, for each $1 \leq i < j \leq k$, the number of edges from V_i to V_j is equal. We can make this assumption without loss of generality because we can easily transform an instance which does not have this property to one which does without changing the number of multicolour cliques; note that if the input does not already satisfy this condition then $k \geq 3$. We set $q := \max\{e(V_i, V_j) : 1 \leq i < j \leq k\}$ (where $e(A, B)$ denotes the number of edges with one endpoint in A and the other in B), and for any pair of sets (V_i, V_j) where $e(V_i, V_j) = q' < q$, we add vertices $\{u_1, \ldots, u_{q-q'}\}$ to V_i and $\{w_1, \ldots, w_{q-q'}\}$ to V_j, and the set of edges $\{u_r w_r : 1 \leq r \leq q - q'\}$; note that the largest cliques created by this process contain two vertices.

We in fact reduce **p-#**Multicolour Clique to a multicolour version of #Given-Size $\mathcal{L}$-Free Subset, defined as follows.

p-#Multicolour $\mathcal{L}$-Free Subset
Input: A k-tuple of disjoint subsets $A_1, \ldots, A_k \subseteq \mathbb{Z}$.
Parameter: k
Question: How many $\mathcal{L}$-free subsets of $A = \bigcup_{1 \leq i \leq k} A_i$ contain exactly one element from each set $A_1, \ldots, A_k$?

It is easy to give an fpt-Turing reduction from **p-#**Multicolour $\mathcal{L}$-Free Subset to #Given-Size $\mathcal{L}$-Free Subset parameterised by k.

Lemma 21. *Let $\mathcal{L}$ be a linear equation. Then* **p-#**Multicolour $\mathcal{L}$-Free Subset $\leq_{\mathrm{T}}^{\mathrm{fpt}}$ #Given-Size $\mathcal{L}$-Free Subset *(where* #Given-Size $\mathcal{L}$-Free Subset *is parameterised by k).*

Proof. Let $(A_1, \ldots, A_k)$ be the input to an instance of **p-#**Multicolour $\mathcal{L}$-Free Subset. For each non-empty $I \subseteq [k]$, we can use our oracle to $\mathcal{L}$-Free Subset to find N_I, the number of $\mathcal{L}$-free subsets of cardinality exactly k in the set $\bigcup_{i \in I} A_i$. This requires $\Theta(2^k)$ oracle calls, and for

each oracle call the parameter value is the same as for the original problem. We can now use an inclusion-exclusion method to compute the number of $\mathcal{L}$-free subsets of size k in A that contain exactly one number from each of the sets A_i: this is precisely

$$\sum_{\emptyset \neq I \subseteq [k]} (-1)^{k-|I|} N_I.$$

□

The main work in the reduction is in the next lemma, where we show that **p-#**Multicolour Clique can be reduced to **p-#**Multicolour $\mathcal{L}$-Free Subset for certain equations $\mathcal{L}$.

Lemma 22. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + a_2x_2 = by$, where $a_1, a_2, b \in \mathbb{N}$ are fixed. Then* **p-#**Multicolour Clique $\leq_{\mathrm{T}}^{\mathrm{fpt}}$ **p-#**Multicolour $\mathcal{L}$-Free Subset.

Proof. Let $(G, \{V_1, \ldots, V_k\})$ be the input to an instance of **p-#**Multicolour Clique, where $G = (V, E)$, and for each $1 \leq i < j \leq k$ let $E_{i,j}$ denote the set of edges between V_i and V_j. We may assume that $|E_{i,j}| = q$ for each $1 \leq i < j \leq k$ and that each V_i is an independent set.

Suppose that $V = \{v_1, \ldots, v_n\}$. We begin by constructing a set $A \subseteq \mathbb{N}$ as in Lemma 1; note that $|A| = \mathcal{O}(|V|^2)$ and $\log(\max(A)) = \mathcal{O}(|V|)$, so $|A| \log(\max(A)) = \mathcal{O}(|V|^3)$. We partition A' into k subsets $A_1, \ldots, A_k$, where $A_i := \phi_V^{-1}(V_i)$, and A'' into $\binom{k}{2}$ subsets $A_{i,j}$ (for $1 \leq i < j \leq k$) where $A_{i,j} := \phi_E^{-1}(E_{i,j})$.

Now let $t \in \{1, \ldots, \binom{k}{2}\}$. We define $X_t \subseteq \mathbb{N}$ to be a set of $t\binom{k}{2} < k^4$ natural numbers disjoint from A, chosen so that every solution to $\mathcal{L}$ in $A \cup X_t$ is contained in A. Without loss of generality, we may assume that $k^4 < |V|$ (otherwise we would be able to execute a brute force approach in time bounded by a function of k alone), so it follows from Lemmas 15 and 17 that we can construct such a set X_t in time bounded by a polynomial function of $|A| \log(\max(A))$, and hence by a polynomial function of $|V|$. Moreover, the space required to represent X_t is also bounded by a polynomial function of $|V|$. We now partition X_t arbitrarily into $\binom{k}{2}$ sets $X_t^{i,j}$ for $1 \leq i < j \leq k$, each of size exactly t. The set $A_{i,j}[t]$ is then defined to be $A_{i,j} \cup X_t^{i,j}$. We set

$$A[t] := \bigcup_{1 \leq i \leq k} A_i \cup \bigcup_{1 \leq i < j \leq k} A_{i,j}[t].$$

It follows from Lemma 1 and the construction of $A[t]$ that the only solutions to $\mathcal{L}$ in $A[t]$ are of the form $a_1x_1 + a_2x_2 = by$, where y corresponds to an edge whose endpoints are the vertices corresponding to x_1 and x_2. We will say that a subset of $A[t]$ is *colourful* if it contains precisely one element from each set A_i (for $1 \leq i \leq k$) and one element from each set $A_{i,j}[t]$ (for $1 \leq i < j \leq k$).

Let $N(A[t])$ denote the number of $\mathcal{L}$-free subsets of $A[t]$ that are colourful. We can compute the value of $N(A[t])$ using a single call to our oracle for **p-#**Multicolour $\mathcal{L}$-Free Subset with input $(A_1, \ldots, A_k, A_{1,2}[t], \ldots, A_{k-1,k}[t])$; note that the total size of the instance in such an oracle call is bounded by $h(k) \cdot |V|^{\mathcal{O}(1)}$ for some function h, and that the value of the parameter in our oracle call depends only on k.

Given any subset $U \subseteq V$ such that $|U \cap V_i| = 1$ for each $1 \leq i \leq k$, let us denote by $N(A[t], U)$ the number of colourful $\mathcal{L}$-free subsets of $A[t]$ whose intersection with A' is precisely $\phi_V^{-1}(U)$. We now claim that

$$N(A[t], U) = (q + t)^{\binom{k}{2} - e(U)} (q + t - 1)^{e(U)},$$

where $e(U)$ denotes the number of edges in the subgraph of G induced by U. To see that this is true, suppose that $U = \{w_1, \ldots, w_k\}$, where $w_i \in V_i$ for each i. If w_i and w_j are not adjacent, then we can choose freely any element of $A_{i,j}[t]$ to add to the set, without risk of creating a solution to $\mathcal{L}$, so there are $|A_{i,j}[t]| = q + t$ possibilities for the element of $A_{i,j}[t]$ in the set; on the other hand, if w_i and w_j are adjacent, we must avoid the element of $A_{i,j}[t]$ corresponding to w_iw_j, so there are

$|A_{i,j}[t]| - 1 = q + t - 1$ possibilities for the element of $A_{i,j}[t]$ in the set. Since we can choose each element of A'' to include in the set independently of the others, this gives the expression above for $N(A[t], U)$.

Observe now that

$$\begin{aligned} N(A[t]) &= \sum_{\substack{U \subseteq V \\ |U \cap V_i| = 1 \text{ for each } 1 \leq i \leq k}} N(A[t], U) \\ &= \sum_{\substack{U \subseteq V \\ |U \cap V_i| = 1 \text{ for each } 1 \leq i \leq k}} (q+t)^{\binom{k}{2} - e(U)} (q+t-1)^{e(U)}. \end{aligned}$$

For $0 \leq j \leq \binom{k}{2}$, let C_j denote the number of subsets $U \subset V$ such that $|U \cap V_i| = 1$ for each $1 \leq i \leq k$ and $e(U) = j$. We can then rewrite the expression above as

$$N(A[t]) = \sum_{j=0}^{\binom{k}{2}} C_j (q+t)^{\binom{k}{2} - j} (q+t-1)^j.$$

If we now define

$$p(z) = \sum_{j=0}^{\binom{k}{2}} C_j z^{\binom{k}{2} - j} (z-1)^j,$$

it is clear that p is a polynomial in z of degree $\binom{k}{2}$, and moreover that $p(z) = N(A[z-q])$. Thus if we know the value of $p(z)$ for $\binom{k}{2} + 1$ distinct values of z, we can interpolate in polynomial time to determine all the coefficients of p; we can obtain the required values of $p(z)$ by using our oracle to evaluate $N(A[t])$ for $t \in \{0, 1, \ldots, \binom{k}{2}\}$.

To complete the reduction we must demonstrate that, once we know the coefficients of $p(z)$, it is straightforward to calculate the number of multicolour cliques in G. We will in fact argue that we only need to determine the constant term of $p(z)$. Note that, if $0 \leq j < \binom{k}{2}$, then z is a factor of $C_j z^{\binom{k}{2} - j} (z-1)^j$. Thus the constant term of $p(z)$ is the same as the constant term of the polynomial

$$C_{\binom{k}{2}} z^{\binom{k}{2} - \binom{k}{2}} (z-1)^{\binom{k}{2}} = C_{\binom{k}{2}} (z-1)^{\binom{k}{2}}.$$

This constant term is

$$C_{\binom{k}{2}} (-1)^{\binom{k}{2}},$$

so the absolute value of the constant term in $p(z)$ is precisely $C_{\binom{k}{2}}$. But $C_{\binom{k}{2}}$ is by definition the number of subsets $U \subseteq V$ such that $|U \cap V_i| = 1$ for each i and $e(U) = \binom{k}{2}$, that is the number of multicolour cliques in G.

This completes the fpt-Turing reduction from **p-#**Multicolour Clique to **p-#**Multicolour $\mathcal{L}$-Free Subset. □

The main result of this section now follows immediately from Lemmas 21 and 22.

Theorem 23. *Let $\mathcal{L}$ be a linear equation of the form $a_1 x_1 + a_2 x_2 = by$ where $a_1, a_2, b \in \mathbb{N}$ are fixed. Then #*Given-Size $\mathcal{L}$-Free Subset*, parameterised by k, is #W[1]-hard with respect to fpt-Turing reductions.*

7.4. **Approximate counting.** For 3-variable homogeneous linear equations $\mathcal{L}$, we have seen that there is unlikely to be an fpt-algorithm, with parameter k, to solve #Given-Size $\mathcal{L}$-Free Subset *exactly*; however, the problem does admit an FPTRAS (for any non-translation-invariant homogeneous linear equation $\mathcal{L}$). We prove the following result, following the method of [35, Lemma 3.4].

Lemma 24. *Let $\mathcal{L}$ be any non-translation-invariant homogeneous linear equation. Let $A \subset \mathbb{Z}$ be finite and $k \in \mathbb{N}$, and let N denote the number of elements of $A^{(k)}$ that are $\mathcal{L}$-free. Then, for every $\varepsilon > 0$ and $\delta \in (0,1)$ there is an explicit randomised algorithm which outputs an integer α, such that*

$$\mathbb{P}[|\alpha - N| \leq \varepsilon \cdot N] \geq 1 - \delta,$$

and runs in time at most $f(k)q(size(A), \varepsilon^{-1}, \log(\delta^{-1}))$, where f is a computable function, q is a polynomial.

Sketch proof. The algorithm uses a simple randomised sampling method: we repeatedly select a random subset of A of cardinality k, and check whether it is $\mathcal{L}$-free. After completing a suitable number of trials, we conclude that the observed proportion of $\mathcal{L}$-free subsets is a good estimate for the overall proportion of such subsets. In order for this conclusion to be valid (without requiring a prohibitively large number of trials) we need to know that the proportion of $\mathcal{L}$-free subsets of size k in A is reasonably large (at least $\dfrac{1}{f(k)\tilde{q}(\text{size}(A))}$, where f is any computable function and $\tilde{q}$ is any polynomial). This follows immediately (if $|A|$ is sufficiently large compared with k; otherwise we can count exactly by brute force in time bounded by a function of k) from the fact that, by Theorem 11 A must contain a large $\mathcal{L}$-free subset B, and any subset of B of size k will necessarily be $\mathcal{L}$-free. □

The existence of an FPTRAS now follows immediately.

Theorem 25. *Let $\mathcal{L}$ be any non-translation-invariant homogeneous linear equation. Then* #Given-Size $\mathcal{L}$-Free Subset*admits an FPTRAS.*

8. An extension version of the problem

A natural variant of the problem $\mathcal{L}$-Free Subset is to ask whether, given $A \subseteq \mathbb{Z}$ and an ($\mathcal{L}$-free) subset $B \subset A$, there is an $\mathcal{L}$-free subset of A of cardinality k which contains B. This problem can be stated formally as follows.

$\mathcal{L}$-Free Subset Extension
Input: A finite set $A \subseteq \mathbb{Z}$, a set $B \subset A$ and $k \in \mathbb{N}$.
Question: Does there exist an $\mathcal{L}$-free subset $A' \subseteq A$ such that $B \subseteq A'$ and $|A'| = k$?

We can make certain easy deductions about the complexity of this problem from the results we have already proved about $\mathcal{L}$-Free Subset. Notice that we can easily define a reduction from $\mathcal{L}$-Free Subset to $\mathcal{L}$-Free Subset Extension by setting $B = \emptyset$; the next result follows immediately from this observation together with Theorem 4.

Proposition 26. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + \cdots + a_\ell x_\ell = by$ where each $a_i \in \mathbb{N}$ and $b \in \mathbb{N}$ are fixed and $\ell \geq 2$. Then $\mathcal{L}$-Free Subset Extension is* NP*-complete, and the problem is para-*NP*-complete parameterised by $|B|$.*

Perhaps the most obvious parameterisation of this problem to consider is $k - |B|$, the number of elements we want to add to the set B. It is straightforward to adapt our earlier results to demonstrate that, in the case of three-term equations, the problem is unlikely to admit an fpt-algorithm with respect to this parameterisation.

Proposition 27. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + a_2x_2 = by$, where $a_1, a_2, b \in \mathbb{N}$ are fixed. Then $\mathcal{L}$-Free Subset Extension is W[1]-hard, parameterised by $k - |B|$.*

Proof. We prove this result by means of a reduction from the W[1]-complete problem **p-**Independent Set, which is defined as follows.

p-Independent Set
Input: A graph G and $k \in \mathbb{N}$.
Parameter: k
Question: Does G contain an independent set of cardinality k?

The W[1]-completeness of this problem can easily be deduced from that of **p-**Clique, shown to be W[1]-complete in [16].

Let (G, k') be the input to an instance of **p-**Independent Set. Once again, we rely on the construction in Lemma 1 to give us the set A in our instance of $\mathcal{L}$-Free Subset Extension; we set $B := A''$ and $k := |B| + k'$ (so the parameter of interest in our instance of $\mathcal{L}$-Free Subset Extension is equal to the solution size in the instance of **p-**Independent Set). By Corollary 2, we know that there is a one-to-one correspondence between independent sets in G of cardinality k' and $\mathcal{L}$-free subsets of A of cardinality $|A''| + k'$ that contain A'', so it follows immediately that (A, B, k) is a yes-instance for $\mathcal{L}$-Free Subset Extension if and only if (G, k') is a yes-instance for **p-**Independent Set. □

On the positive side, we observe that we can once again make use of fpt-algorithms for **p**-card-Hitting Set if we consider $\mathcal{L}$-Free Subset Extension parameterised by the number of elements of A that are *not* included in the subset. Note that the standard bounded search tree method for **p**-card-Hitting Set (or its counting version) [23, Theorem 1.14] in fact gives an fpt-algorithm to find all hitting sets of size k. As it is easy to check in time polynomial in size(A) whether a given hitting set in the hypergraph defined in the proof of Lemma 13 contains a vertex corresponding to an element of B, we can use this method to count $\mathcal{L}$-free subsets of A that contain B (and hence to decide whether there is at least one).

Proposition 28. *Let $\mathcal{L}$ be any fixed linear equation. Then $\mathcal{L}$-Free Subset Extension, parameterised by $|A| - k$, belongs to* FPT*; the same is true for the counting version of the problem with this parameterisation.*

Finally, we consider parameterising simultaneously by the number of elements we wish to add to B and the size of the set B; this is equivalent to parameterising by k, the total size of the desired $\mathcal{L}$-free subset.

Proposition 29. *Let $\mathcal{L}$ be a linear equation of the form $a_1x_1 + a_2x_2 = by$, where $a_1, a_2, b \in \mathbb{N}$ are fixed and $a_1 + a_2 \neq b$. Then $\mathcal{L}$-Free Subset Extension, parameterised by k, belongs to* FPT*.*

Proof. Let (A, B, k) be an instance of $\mathcal{L}$-Free Subset Extension. If $k < |B|$ then this is necessarily a no-instance, so we may assume without loss of generality that $k \geq |B|$; we may also assume that B is $\mathcal{L}$-free (we can check this in polynomial time and if B contains a solution to $\mathcal{L}$ we immediately return NO).

As a first step in our algorithm, we delete from A every $a \in A \setminus B$ such that $B \cup \{a\}$ is not $\mathcal{L}$-free: note that this does not change the number of $\mathcal{L}$-free subsets containing B, as such an a cannot belong to any set of this kind. We call the resulting set A_1, and note that we can construct A_1 in time polynomial in size(A). Note that no set containing 0 can be $\mathcal{L}$-free, so we know that $0 \notin A_1$.

Fix the constant $\lambda = \lambda(\mathcal{L})$ as in equation (2). Our algorithm proceeds as follows. If $|A_1| < \left(\frac{6|B|+1}{\lambda}\right)(k-|B|)+|B|$, we exhaustively consider all k-element subsets of A_1 and check if they form an $\mathcal{L}$-free subset; if $|A_1| \geq \left(\frac{6|B|+1}{\lambda}\right)(k-|B|)+|B|$ then we return YES. To see that this is an fpt-algorithm, note that, if $|A_1| < \left(\frac{6|B|+1}{\lambda}\right)(k-|B|)+|B|$, then $|A_1| = \mathcal{O}(k^2)$, so we can perform the exhaustive search in time depending only on k. It is clear that we will return the correct answer whenever we perform the exhaustive search; in order to prove correctness of the algorithm, it remains to show that, if $|A_1| \geq \left(\frac{6|B|+1}{\lambda}\right)(k-|B|)+|B|$, then we must have a yes-instance.

To see that this is true, we first prove that there exists a large set $A_2 \subseteq A_1$ such that $B \subseteq A_2$ and no solution to $\mathcal{L}$ in A_2 involves an element of B. We will call a triple $(x,y,z) \in A_1^3$ *bad* if $a_1x + a_2y = bz$. By construction of A_1, note that every bad triple contains at most one element of B. We aim to bound the number of bad triples containing some fixed $u \in A_1 \setminus B$ and at least one element of B. First, we also fix $v \in B$, and bound the number of bad triples which contain both u and v. If we fix the positions of u and v in a triple, there is at most one $w \in A_1 \setminus B$ such that w completes the triple; as there are 6 options for the choice of positions of u and v in the triple, this means there are in total at most 6 bad triples involving both u and v. Summing over all possibilities for v, we see that there are at most $6|B|$ bad triples involving any fixed $u \in A_1 \setminus B$ and at least one element of B.

We can therefore greedily construct a set $C \subseteq A_1 \setminus B$ of size at least $\frac{|A_1|-|B|}{6|B|+1}$ such that every bad triple in $B \cup C$ is entirely contained in C. Indeed, initially set $C := \emptyset$ and $A' := A_1 \setminus B$. Move an arbitrary element u of A' to C and delete all elements of A' that form a bad triple with u and at least one element of B; by the reasoning above, this involves deleting at most $6|B|$ elements of A'. Repeat this process until $A' = \emptyset$, and note that C is as desired; set $A_2 := B \cup C$.

Now observe that every $\mathcal{L}$-free subset of cardinality $k-|B|$ in C can be extended to an $\mathcal{L}$-free subset of A_2 of cardinality k which contains B. We know from Theorem 11 that there exists an $\mathcal{L}$-free subset $C' \subseteq C$ of cardinality at least $\lambda|C|$ (where λ is the constant defined in equation (2)). $B \cup C'$ is then an $\mathcal{L}$-free subset of A, and has cardinality at least

$$|B| + \lambda\left(\frac{|A_1|-|B|}{6|B|+1}\right) \geq |B| + \lambda\left(\frac{\left(\frac{6|B|+1}{\lambda}\right)(k-|B|)+|B|-|B|}{6|B|+1}\right) = k,$$

so we should indeed return YES. □

We note that the argument used in this proof can be adapted to demonstrate the existence of an FPTRAS for the counting version of this problem, using the ideas from Lemma 24. However, there is unlikely to be an fpt-algorithm to solve the counting version exactly, as we can easily reduce #Given-Size $\mathcal{L}$-Free Subset to this problem (with the same parameter) by setting $B = \emptyset$.

9. Conclusions and open problems

We have shown that the basic problem of deciding whether a given input set $A \subseteq \mathbb{Z}$ contains an $\mathcal{L}$-free subset of size at least k is NP-complete when $\mathcal{L}$ is any linear equation of the form

$a_1x_1 + \cdots + a_\ell x_\ell = by$ (with $a_i, b \in \mathbb{N}$ and $\ell \geq 2$), although the problem is solvable in polynomial time whenever $\mathcal{L}$ is a linear equation with only two variables. We also demonstrated that the maximisation version of the problem is APX-hard for equations $\mathcal{L}$ of the form $a_1x_1 + a_2x_2 = by$ (with $a_1, a_2, b \in \mathbb{N}$).

Two natural questions arise from these results. First of all, in our NP-hardness reduction, we construct a set A where $\max(A)$ is exponential in terms of $|A|$: is this problem in fact *strongly* NP-complete, so that it remains hard even if all elements of A are bounded by some polynomial function of $|A|$? Secondly, can either the NP-completeness proof or the APX-hardness proof be generalised to other linear equations $\mathcal{L}$? A natural starting point for an equation that is not covered by Theorem 4 would perhaps be the case of Sidon sets (i.e. $x + y = z + w$).

On the positive side, we saw that the decision problem belongs to FPT for any homogeneous non-translation-invariant equation $\mathcal{L}$ when parameterised by the cardinality of the desired $\mathcal{L}$-free subset, and that it belongs to FPT for any linear equation $\mathcal{L}$ with respect to the dual parameterisation (the number of elements of A not included in the $\mathcal{L}$-free subset). While we have considered two natural parameterisations here, there is another natural parameterisation that we have not considered. We know that, for certain linear equations, there is some function $c^*_{\mathcal{L}}$ such that every set $A \subseteq \mathbb{Z}$ is certain to contain an $\mathcal{L}$-free subset of cardinality at least $c^*_{\mathcal{L}}(|A|)$. It is therefore natural to consider the complexity parameterised above this lower-bound: what is the complexity of determining whether a given subset $A \subseteq \mathbb{Z}$ contains an $\mathcal{L}$-free subset of cardinality at least $c^*_{\mathcal{L}}(|A|) + k$, where k is taken to be the parameter? The main difficulty in addressing this question is that the exact value of $c^*_{\mathcal{L}}$ is not known for any linear equations $\mathcal{L}$: even in the case of sum-free subsets, we only know that the bound on $c^*_{\mathcal{L}}$ is of the form $\frac{|A|}{3} + o(|A|)$.

We also considered the complexity of determining whether a set $A \subseteq \mathbb{Z} \setminus \{0\}$ contains an $\mathcal{L}$-free subset containing a fixed proportion ε of its elements. We demonstrated that this problem is also NP-complete for the case of sum-free sets, and also for $\mathcal{L}$-free sets whenever $\mathcal{L}$ is a 3-variable translation-invariant linear equation. It would be interesting to investigate how far these results can be generalised to other linear equations: given any non-translation-invariant, homogeneous linear equation $\mathcal{L}$ and any rational $\kappa(\mathcal{L}) < \varepsilon < 1$, is ε-$\mathcal{L}$-Free Subset NP-complete?

Concerning the complexity of counting $\mathcal{L}$-free subsets, we have addressed the problem of counting $\mathcal{L}$-free subsets containing exactly k elements. For equations $\mathcal{L}$ covered by the NP-hardness result of Theorem 4, even approximate counting is hard: there is no FPRAS for arbitrary k unless NP = RP (as if we could count approximately we could, with high probability, determine whether there is at least one $\mathcal{L}$-free subset of size k). We also considered the complexity of this problem parameterised separately by k and $|A| - k$.

However, there are other natural counting problems we have not addressed here. For example, we might want to count the total number of $\mathcal{L}$-free subsets of *any* size; here the decision problem ("Is there an $\mathcal{L}$-free subset of any size?") is trivial, so there is no immediate barrier to an efficient counting algorithm. Alternatively, we might want to count the total number of *maximal* $\mathcal{L}$-free sets. Our results do not have any immediate consequences for either of these problems, but the corresponding counting versions of the extension problem are necessarily #P-complete: by Corollary 2, we have a polynomial-time reduction to this problem from that of counting all (maximal) independent sets in an arbitrary ℓ-uniform hypergraph; for $\ell = 2$ and $\ell = 3$ this problem was shown to be #P-complete by Greenhill [30], and we can easily add further dummy vertices to each edge (and then require that the elements corresponding to the dummy vertices are included in our $\mathcal{L}$-free subset) to deal with larger values of ℓ.

Acknowledgment

The authors are grateful to the referees for their enthusiastic and careful reviews.

References

[1] Manindra Agrawal, Neeraj Kayal and Nitin Saxena, PRIMES is in P, *Ann. of Math.* (2004) 160(2): 781-793.

[2] Paola Alimonti and Viggo Kann, Some APX-completeness results for cubic graphs, *Theor. Comput. Sci.* (2000) 237(1-2): 123-134.

[3] N. Alon and D. J. Kleitman, Sum-free subsets, in *A Tribute to Paul Erdős*, Cambridge Univ. Press, Cambridge, 1990, 13–26.

[4] N. Alon and J.H. Spencer, The Probabilistic Method, John Wiley & Sons, 2004

[5] L. Babai, V. Sós, Sidon sets in groups and induced subgraphs of Cayley graphs, *European J. Combin.*, 6, (1985), 101–114.

[6] J. Balogh, H. Liu, M. Sharifzadeh, The number of subsets of integers with no k-term arithmetic progression, *Int. Math. Res. Not.*, to appear.

[7] J. Balogh, H. Liu, M. Sharifzadeh and A. Treglown, The number of maximal sum-free subsets of integers, *Proc. Amer. Math. Soc.*, 143, (2015), 4713–4721.

[8] J. Balogh, H. Liu, M. Sharifzadeh and A. Treglown, Sharp bound on the number of maximal sum-free subsets of integers, submitted.

[9] J. Balogh, R. Morris and W. Samotij, Independent sets in hypergraphs, *J. Amer. Math. Soc.* **28** (2015), 669–709.

[10] T.F. Bloom, A quantitative improvement for Roth's theorem on arithmetic progressions, *J. London Math. Soc.*, **93** (3), (2016), 643–663.

[11] J. Bourgain, Estimates related to sumfree subsets of sets of integers, *Israel J. Math.*, 97, (1997), 71–92.

[12] T.C. Brown and J.C. Buhler, A density version of a geometric Ramsey theorem, *J. Combin. Theory, Ser. A* **32** (1982), 20–34.

[13] P. Cameron and P. Erdős, On the number of sets of integers with various properties, in Number Theory (R.A. Mollin, ed.), 61–79, Walter de Gruyter, Berlin, 1990.

[14] P. Cameron and P. Erdős, Notes on sum-free and related sets, *Combin. Probab. Comput.*, 8, (1999), 95–107.

[15] P.H. Diananda and H.P. Yap, Maximal sum-free sets of elements of finite groups, *Proc. Japan Acad.*, **45**, (1969), 1–5.

[16] R.G. Downey and M.R. Fellows. Fixed-parameter tractability and completeness II: On completeness for W[1]. Theoretical Computer Science, 141:109131, 1995.

[17] Rodney G. Downey and Michael R. Fellows, *Fundamentals of Parameterized Complexity*, Springer London, 2013.

[18] S. Eberhard, B. Green and F. Manners, Sets of integers with no large sum-free subset, *Ann. Math.*, 180, (2014), 621–652.

[19] M. Elkin, An improved construction of progression-free sets, *Israel J. Math.*, 184, (2011), 93–128.

[20] P. Erdős, Extremal problems in number theory, *Proceedings of the Symp. Pure Math.* VIII AMS (1965), 181–189.

[21] M. Fellows, D. Hermelin, F. Rosamond, and S. Vialette, *On the parameterized complexity of multiple-interval graph problems*, Theoretical Computer Science **410** (2009), 53–61.

[22] J. Flum and M. Grohe, *The parameterized complexity of counting problems*, SIAM Journal on Computing **33** (2004), no. 4, 892–922.

[23] J. Flum and M. Grohe, *Parameterized complexity theory*, Springer, 2006.

[24] P. Frankl, G Graham, and V. Rödl, On subsets of abelian groups with no 3-term arithmetic progression, *J. Combin. Theory, Ser. A* **45** (1987), 157–161.

[25] Vincent Froese, Iyad Kanj, André Nichterlein and Rolf Niedermeier, Finding Points in General Position, In Proceedings of the 28th Canadian Conference on Computational Geometry (CCCG '16), pages 714. 2016.

[26] M.R. Garey and D.S. Johnson, Computers and intractability, Freeman, 1979.

[27] B. Green, The Cameron-Erdős conjecture, *Bull. London Math. Soc.*, 36, (2004), 769–778.

[28] B. Green and I. Ruzsa, Sum-free sets in abelian groups, *Israel J. Math.*, 147, (2005), 157–189.

[29] B. Green and J. Wolf, A note on Elkins improvement of Behrend's construction, *Additive number theory: Festschrift in honor of the sixtieth birthday of Melvyn B. Nathanson*, pages 141–144. Springer-Verlag, 1st edition, 2010.

[30] C. Greenhill, The complexity of counting colourings and independent sets in sparse graphs and hypergraphs, *Comput. complex.* (2000) 9:52, doi:10.1007/PL00001601.

[31] W.T. Gowers, Quasirandom groups, *Combin. Probab. Comput.*, 17, (2008), 363–387.

[32] R. Hancock and A. Treglown, On solution-free sets of integers II, *Acta Arith.*, 180, (2017), 15–33.

[33] M.N. Kolountzakis, Selection of a large sum-free subset in polynomial time, *Inform. Process. Lett.* **49** (1994), 255–256.

[34] V.F. Lev, Progression-free sets in finite abelian groups, *J. Number Theory* **104** (2004), 162–169.

[35] Kitty Meeks, The challenges of unbounded treewidth in parameterised subgraph counting problems, *Discrete Applied Mathematics* **198** (2016), pp.170-194; doi:10.1016/j.dam.2015.06.019.

[36] K.F. Roth, On certain sets of integers, *J. London Math. Soc.*, 28, (1953), 104–109.
[37] I.Z. Ruzsa, Solving a linear equation in a set of integers I, *Acta Arith.*, 65, (1993), 259–282.
[38] I.Z. Ruzsa, Solving a linear equation in a set of integers II, *Acta Arith.*, 72, (1995), 385–397.
[39] T. Sanders, On Roth's theorem on progressions, *Ann. of Math.*, 174, (2011), 619–636.
[40] A.A. Sapozhenko, The Cameron-Erdős conjecture, (Russian) *Dokl. Akad. Nauk.*, 393, (2003), 749–752.
[41] D. Saxton and A. Thomason, Hypergraph containers, *Invent. Math.* **201** (2015), 925–992.
[42] I. Schur, Uber die Kongruenz $x^m + y^m \equiv z^m \pmod{p}$, *ber. Deutsch. Mat. Verein.*, 25, (1916), 114–117.
[43] M. Thurley, Kernelizations for Parameterized Counting Problems, In: Cai JY., Cooper S.B., Zhu H. (eds) Theory and Applications of Models of Computation. TAMC 2007. Lecture Notes in Computer Science, vol 4484. Springer, Berlin, Heidelberg.
[44] E. Szemerédi, On sets of integers containing no k elements in arithmetic progression, *Acta Arith.*, **27** (1975), 199–245.
[45] T. Tao and V. Vu, Sumfree sets in groups: a survey, submitted.
[46] B.L. van der Wareden, Beweis einer Baudetschen Vermutung, *Nieuw Arch. Wisk.* **15** (1927), 212–216.

Kitty Meeks
School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
Glasgow
G12 8QQ
UK

Andrew Treglown
School of Mathematics
University of Birmingham
Edgbaston
Birmingham
B15 2TT
UK

E-mail addresses: kitty.meeks@glasgow.ac.uk, a.c.treglown@bham.ac.uk