The complexity of solution-free sets of integers for general linear equations

Edwards, Keith; Noble, Steven D.

Published in:
Discrete Applied Mathematics

DOI:
10.1016/j.dam.2019.07.008

Publication date:
2019

The complexity of solution-free sets of integers for general linear equations

Keith J. Edwards*
Computing
University of Dundee
Dundee, DD1 4HN
United Kingdom
kjedwards@dundee.ac.uk

Steven D. Noble
Department of Economics, Mathematics and Statistics
Birkbeck, University of London
Malet Street
London, WC1E 7HX
United Kingdom
s.noble@bbk.ac.uk

June 24, 2019

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 $\mathcal{L}$. Meeks and Treglown [6] showed that for certain kinds of linear equations, it is NP-complete to decide if a given set of integers contains a solution-free subset of a given size. Also, for equations involving three variables, they showed that the problem of determining the size of the largest solution-free subset is APX-hard, and that for two such equations (representing sum-free and progression-free sets), the problem of deciding if there is a solution-free subset with at least a specified proportion of the elements is also NP-complete.

We answer a number of questions posed by Meeks and Treglown, by extending the results above to all linear equations, and showing that the problems remain hard for sets of integers whose elements are polynomially bounded in the size of the set. For most of these results, the integers can all be positive as long as the coefficients do not all have the same sign.

We also consider the problem of counting the number of solution-free subsets of a given set, and show that this problem is #P-complete for any linear equation in at least three variables.

Keywords: solution-free set; computational complexity.

*Corresponding author

1 Introduction

There has been much study of sum-free or progression-free sets of integers, that is, sets of integers containing no solution to the equations $x + y = z$ or $x + z = 2y$ respectively, or, more generally, sets containing no solution to some other linear equation.

Most of this research has focussed on the existence or otherwise of large sum-free subsets of an arbitrary set, rather than complexity questions. For instance Erdős [4] proved that every set of n non-zero integers contains a sum-free subset of size at least $n/3$. In contrast it follows from Roth's theorem [8] that the largest progression-free subset of $\{1, \ldots, n\}$ has size $o(n)$.

Many more results on sum-free and progression-free sets are mentioned in the introduction to the recent paper by Meeks and Treglown [6], which initiated the study of the computational complexity aspects of these problems, by considering several computational problems concerning solution-free sets, or, more precisely, solution-free subsets of a given set of integers.

1.1 Solution-free sets

We start with a brief account of the notation and terminology that will be used throughout the paper.

Consider a fixed linear equation $\mathcal{L}$ of the form

$$c_1 x_1 + \cdots + c_\ell x_\ell = K, \tag{1}$$

where $c_1, \ldots, c_\ell, K \in \mathbb{Z}$. The equation $\mathcal{L}$ is *homogeneous* if $K = 0$ and inhomogeneous otherwise. It is *translation-invariant* if

$$\sum_{i=1}^{\ell} c_i = 0.$$

We will be interested in solutions to such an equation in some set of integers. In general, a solution to $\mathcal{L}$ in a set $A \subseteq \mathbb{Z}$ is a sequence $(x_1, \ldots, x_\ell) \in A^\ell$ which satisfies the equation.

However, some equations have trivial solutions which must be excluded from consideration in order to obtain sensible questions. Suppose that $\mathcal{L}$ is homogeneous and translation-invariant, that is, it is of the form

$$c_1 x_1 + \ldots + c_\ell x_\ell = 0, \tag{2}$$

where $c_1 + \ldots + c_\ell = 0$. Then $(x, \ldots, x)$ is a trivial solution of (2) for any x (and so no non-empty set of integers can be solution-free for $\mathcal{L}$). More generally, a solution $(x_1, \ldots, x_\ell)$ to $\mathcal{L}$ is said to be *trivial* if there exists a partition $P_1, \ldots, P_k$ of $\{1, \ldots, \ell\}$ so that:

(i) $x_i = x_j$ for every i, j in the same partition class P_r;

(ii) For each $r \in \{1, \ldots, k\}$, $\sum_{i \in P_r} c_i = 0$.

For any linear equation $\mathcal{L}$, a set A of integers is *$\mathcal{L}$-free* if A does not contain any non-trivial solutions to $\mathcal{L}$.

2 Computational Problems

Meeks and Treglown [6] considered several computational problems concerning $\mathcal{L}$-free sets for some fixed linear equation $\mathcal{L}$ (they restricted attention to homogeneous equations).

The main problem considered was the following:

> $\mathcal{L}$-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 $|A'| = k$?

They show that the problem is NP-complete in some cases:

Theorem 2.1 (Meeks & Treglown [6]). *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 is NP-complete.*

They also consider the problem of finding the size of the largest $\mathcal{L}$-free subset of a set:

> 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 $A' \subseteq A$?

This problem is shown to be APX-hard in some limited cases with three variables:

Theorem 2.2 (Meeks & Treglown [6]). *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.*

Meeks and Treglown also consider the problem of determining if a set has an $\mathcal{L}$-free subset containing some fixed proportion of the elements:

> 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 $|A'| \geq \varepsilon|A|$?

This problem is shown to be NP-complete for some values of ε for two kinds of equations, namely $x + y = z$ and $ax + by = cz$, where $a, b, c > 0$ and $a + b = c$.

The authors raise a number of questions, including the following:

1. Can these results be proved for the case when the set A has elements of size polynomial in the size of A (rather than exponential as is the case with the results proved in [6])?
2. Can the results be generalised to other linear equations, in particular equations such as $x + y = z + w$?

We answer both these questions in the affirmative, by extending all the above results to all linear equations (both homogeneous and inhomogeneous) with at least three variables, using only polynomially-sized integers.

The structure of the rest of the paper is as follows: In Section 3 we give the first of several

similar representations of a graph as a set of integers, relative to a fixed homogeneous linear equation. We then use this to show (Theorem 3.5) that the problem $\mathcal{L}$-Free Subset is strongly NP-complete for any such equation with at least three variables. We then extend this to inhomogeneous equations using a slightly different construction (Theorem 3.9). In Section 4 we consider the complexity of determining the size of the largest solution-free subset of a set. In Section 5 we consider the problem of deciding if a set has a solution-free subset containing a given proportion of its elements and prove that ε-$\mathcal{L}$-Free Subset is strongly NP-complete for any linear equation with at least three variables (Theorem 5.3). Finally in Section 6 we prove the hardness of the counting version of $\mathcal{L}$-Free Subset (Theorem 6.8).

3 Representing a graph as a set of integers

In this section we will describe constructions which allow us to represent a graph as a set of integers relative to a linear equation. The constructions are inspired by the hypergraph one used in [6], but are different in a number of respects; in particular, we do not use hypergraphs. For an equation with ℓ variables, instead of associating $\ell-1$ variables with vertices and the remaining (dependent) one with an edge of an $(\ell-1)$-uniform hypergraph, we associate two variables with vertices and the remaining $\ell-2$ with an edge, one of which is "dependent" and the rest "free". We will start with homogeneous equations.

We first show how to rearrange a homogeneous equation into a more convenient form. Note that this only involves re-ordering the terms. This form is helpful when showing that the integers in the set can all be positive provided the coefficients do not all have the same sign.

Lemma 3.1. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero. Then we can find an equivalent equation $\mathcal{L}'$: $a_1x_1 + a_2x_2 + b_1y_1 + \cdots + b_{\ell-3}y_{\ell-3} = b_0y_0$, with the following properties:*

1. *$a_1 \leq a_2 \leq b_1 \leq \ldots \leq b_{\ell-3}$.*
2. *$b_0 > 0$ unless all of $c_1, \ldots, c_\ell$ have the same sign, in which case $b_0 < 0$.*
3. *All of $a_1, a_2, b_1, \ldots, b_{\ell-3}$ are non-zero,*
4. *If $C = \{a_1, a_2, b_1, \ldots, b_{\ell-3}\}$, then $\sum_{x\in C, x<0}(-x) < \sum_{x\in C, x>0} x$, that is, the negative coefficients on the left-hand side have smaller total size than the positive coefficents on the left-hand side.*
5. *The last two coefficients on the left-hand side are positive (these are a_1, a_2 if $\ell = 3$, a_2, b_1 if $\ell = 4$, $b_{\ell-4}, b_{\ell-3}$ if $\ell > 4$).*

Proof. First, if all the coefficients have the same sign, then we can assume they are all strictly positive. Order the coefficients in increasing order, and set $a_1 = c_1$, $a_2 = c_2$, $b_i = c_{i+2}$ for $i = 1, \ldots, \ell-3$, and $b_0 = -c_\ell$.

Now suppose that not all the coefficients have the same sign. If there is only one positive coefficient, or only one negative, we can reverse signs if necessary to ensure there is exactly one negative coefficient. Otherwise, by reversing the signs of all the coefficients if necessary, we can assume that (the negation of) the sum of the negative coefficients

is at most the sum of the positive coefficients. In either case, reorder the terms so that the coefficients are in non-decreasing order, then the first term has a negative coefficient since not all coefficients have the same sign; move this term to the right-hand side. Then rename the coefficients and variables to obtain the required form. □

Note: We will say that the equation $\mathcal{L}'$ is in standard form.

We now prove a lemma which shows that a graph may be represented as a set of integers relative to a fixed homogeneous linear equation. In this and later variants, each vertex and each edge is represented by one or more integers; these integers are all distinct.

Since we will often wish to refer to the maximum of the absolute values of the elements of a set, we make the following definition.

Definition. For any non-empty set A of integers, let $\mathrm{maxabs}(A) = \max\{|a| : a \in A\}$.

Lemma 3.2. *Consider a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero, and let $\mathcal{L}$ be an equivalent linear equation $a_1x_1 + a_2x_2 + b_1y_1 + \cdots + b_{\ell-3}y_{\ell-3} = b_0y_0$ in standard form. Let $G = (V, E)$ be a graph, and let $n = |V|$. Then we can construct in polynomial time a set $A \subseteq \mathbb{Z}$ with the following properties:*

1. *A is the union of $\ell - 1$ sets $A_V, A_E^0, A_E^1, \ldots, A_E^{\ell-3}$, where $A_V = \{x_v : v \in V\}$ and $A_E^i = \{y_e^i : e \in E\}$ for each $i = 0, 1, \ldots, \ell - 3$;*
2. *$|A| = |V| + (\ell - 2)|E|$;*
3. *for every edge $e = (v, w)$, some permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ is a solution to $\mathcal{L}$; such a solution we call an edge-solution;*
4. *if $(z_1, \ldots, z_\ell)$ is a non-trivial solution to $\mathcal{L}$, then $(z_1, \ldots, z_\ell)$ is a permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ for some edge $e = (v, w)$;*
5. *$\mathrm{maxabs}(A) = \mathcal{O}(|V|^{2\ell+2})$;*
6. *$A \subseteq \mathbb{N}$ unless all the coefficients $c_1, \ldots, c_\ell$ have the same sign.*

Proof. We need to choose the elements of A to satisfy the Conditions 1–6. The set A consists of a set A_V of vertex labels (one per vertex) and a set $A_E = A_E^0 \cup A_E^1 \cup \ldots \cup A_E^{\ell-3}$ of edges labels ($\ell - 2$ per edge).

In order to choose the labels we will use a set of variables, one for each element of A, and construct a set of constraints (inequalities and equations) which these must satisfy. We then choose values for each variable so that all the constraints are satisfied. We will use upper case for each variable name and the corresponding lower case for its value. The set of variables is $B = B_V \cup B_E$, where $B_E = B_E^0 \cup B_E^1 \cup \cdots \cup B_E^{\ell-3}$, $B_V = \{X_v : v \in V\}$ and $B_E^i = \{Y_e^i : e \in E\}$ for each $i = 0, 1, \ldots, \ell - 3$. The variable X_v is associated with the vertex v, and the variables $Y_e^0, \ldots, Y_e^{\ell-3}$ with the edge e. We will call the elements of B_V vertex variables, those of $B_E^1 \cup \cdots \cup B_E^{\ell-3}$ free edge variables, and those of B_E^0 dependent edge variables.

We now consider the constraints that the variables must satisfy. Fix an arbitrary ordering of the vertex set V.

First, for each edge $e = (v, w)$, we have a set of variables corresponding to the edge and its endpoints, that is $\{X_v, X_w, Y_e^1, \ldots, Y_e^{\ell-3}, Y_e^0\}$. We will call such a set an edge-set of

variables. To ensure that Condition 3 is satisfied, we will require these to satisfy the equation $\mathcal{L}$ in a specific way, such that

$$a_1X_v + a_2X_w + b_1Y_e^1 + \cdots + b_{\ell-3}Y_e^{\ell-3} = b_0Y_e^0, \qquad (*)$$

where $v < w$ in the vertex ordering.

Second, we have a set of inequalities of two types. In each case we require that some linear combination of variables is non-zero. Type 1 inequalities ensure that all the labels are distinct, so that Condition 2 is satisfied, while Type 2 inequalities ensure that there are no non-trivial solutions to the equation $\mathcal{L}$, other than edge-solutions, so that Condition 4 is satisfied.

For any sequence $z = (W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0) \in B^\ell$ of variables, we say that z is an edge-sequence if $\{W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0\}$ is an edge-set.

The inequalities are as follows:

Type 1: For any distinct pair $z = (Z, Z') \in B^2$, we have the linear combination $\mathrm{lc}(z) = Z - Z'$, and require that $\mathrm{lc}(z) \neq 0$.

Type 2: For any sequence $z = (W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0) \in B^\ell$, we have the linear combination $\mathrm{lc}(z) = a_1W_1 + a_2W_2 + b_1Z_1 + \cdots + b_{\ell-3}Z_{\ell-3} - b_0Z_0$ (note that the variables in z do not need to be distinct). Provided that (i) z is not an edge-sequence and (ii) $\mathrm{lc}(z)$ is not identically zero, then we require that $\mathrm{lc}(z) \neq 0$. Note that $\mathrm{lc}(z)$ being identically zero corresponds to a trivial solution of the equation, where the total coefficient of each variable is zero and so the terms cancel.

Now when choosing values to satisfy all of the inequalities, we must take account of the constraints $(*)$. Thus in each of the linear combinations $\mathrm{lc}(z)$, we substitute for each dependent edge variable, that is, for any $e = (v, w)$, set $Y_e^0 = (a_1X_v + a_2X_w + b_1Y_e^1 + \cdots + b_{\ell-3}Y_e^{\ell-3})/b_0$ and collect terms to obtain a reduced (rational) linear combination $\mathrm{rlc}(z)$ of the vertex variables and free edge variables in the set $B_V \cup B_E^1 \cup \cdots \cup B_E^{\ell-3}$. This could potentially cause cancellation of terms; we discuss this issue in detail below.

It is clear that if values are chosen for all the variables, satisfying the constraints $(*)$, then the values of $\mathrm{lc}(z)$ and $\mathrm{rlc}(z)$ will be equal. Hence if each inequality $\mathrm{lc}(z) \neq 0$ is replaced by the corresponding inequality $\mathrm{rlc}(z) \neq 0$ and the values of the variables are chosen so that these new inequalities are satisfied, then the original inequalities, expressed in terms of $\mathrm{lc}(z)$, must also be satisfied. Each inequality $\mathrm{rlc}(z) \neq 0$ can be satisfied providing $\mathrm{rlc}(z)$ is not identically zero. Note however that potentially the reduced combination $\mathrm{rlc}(z)$ is identically zero even though $\mathrm{lc}(z)$ is not, because terms may cancel after substituting for the dependent edge variables Y_e^0. Hence before we choose values for the variables, we must check that this can never occur, by proving the following claim.

Claim. If z is not an edge-sequence and $\mathrm{lc}(z)$ is not identically zero, then $\mathrm{rlc}(z)$ can never be identically zero.

For Type 1, $\mathrm{rlc}(z)$ cannot be identically zero since the variables Z, Z' are distinct, and even if at least one is of the form Y_e^0, they cannot cancel out.

For Type 2, consider any sequence $z = (W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0) \in B^\ell$. Suppose that $\mathrm{rlc}(z)$ is identically zero. Then we need to show that either z is an edge-sequence or $\mathrm{lc}(z)$ is identically zero.

First suppose that $\ell \geq 4$ (so that each edge has at least one free label). If none of $W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0$ is a dependent edge variable, then there is no substitution, so $\mathrm{lc}(z) = \mathrm{rlc}(z)$, hence $\mathrm{lc}(z)$ is identically zero.

So suppose that at least one of $W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0$ is a dependent edge variable. If the total coefficient of each dependent edge variable in $\mathrm{lc}(z)$ is zero, then the other terms are the same in $\mathrm{rlc}(z)$ as in $\mathrm{lc}(z)$, so must reduce to zero, hence $\mathrm{lc}(z)$ is identically zero.

Thus we may assume that z contains some dependent edge variable Y_e^0 whose terms do not cancel, that is, the total coefficient of Y_e^0 in $\mathrm{lc}(z)$ is non-zero. Then all of $Y_e^1, \ldots, Y_e^{\ell-3}$ must also occur in the sequence z, since if Y_e^k does not occur in z, the coefficient of Y_e^k in $\mathrm{rlc}(z)$ is non-zero, contrary to assumption.

Hence at least $\ell - 2$ of the terms in the sequence z are edge variables of e, leaving at most two other terms. If one of them is a (free or dependent) edge variable of some edge $e' \neq e$, then $\mathrm{rlc}(z)$ will have a non-zero coefficient of $Y_{e'}^k$ for some $k \geq 1$ unless the final term is also an edge variable of e'. But then there is an endpoint v of e which is not an endpoint of e', and the variable X_v has non-zero coefficient in $\mathrm{rlc}(z)$, contrary to assumption.

Thus the two remaining terms in z must be vertex variables and so must be the variables of the endpoints of e. But then z is an edge-sequence.

Finally consider the case when $\ell = 3$. In this case $\mathrm{lc}(z)$ is just $a_1W_1 + a_2W_2 - b_0Z_0$, with a_1, a_2 positive, and for each edge $e = (v, w)$, the only edge variable is the dependent one $Y_e^0 = (a_1X_v + a_2X_w)/b_0$. If exactly one of W_1, W_2, Z_0 is an edge variable, then the other two must be the variables of its endpoints, so z is an edge-sequence. If exactly two are edge variables, these two terms cannot cancel in $\mathrm{lc}(z)$, for then the third term would be non-zero in $\mathrm{rlc}(z)$. So either the two terms come from the same edge $e = (v, w)$, so that their sum makes a non-zero contribution to the coefficients of the two vertex variables X_v and X_w in $\mathrm{rlc}(z)$, or they come from distinct edges with at least three endpoints between them, and then the sum of these two terms must have non-zero coefficient of at least two vertex variables. In either case, since the remaining term is a variable of a single vertex, at least one vertex has non-zero coefficient in $\mathrm{rlc}(z)$. Finally, suppose all three are edge variables. If all three variables are from the same edge e, then $\mathrm{lc}(z) = cY_e^0$ for some c, and since $\mathrm{rlc}(z)$ is identically zero, c must be zero, so $\mathrm{lc}(z)$ is also identically zero. If one variable is from edge e, and two from edge e', where $e' \neq e$, then some vertex v is an endpoint of e but not of e', and X_v has non-zero coefficient in $\mathrm{rlc}(z)$, contrary to assumption. So all three variables are from distinct edges. Then either there is a vertex which is an endpoint of just one of them, so that its variable has non-zero coefficient in $\mathrm{rlc}(z)$, or the edges form a triangle, in which case one of the vertices occurs only in W_1 and W_2 and so since a_1 and a_2 are positive its variable has non-zero coefficient in $\mathrm{rlc}(z)$.

We conclude that if $\mathrm{rlc}(z)$ is identically zero, then either z is an edge-sequence or $\mathrm{lc}(z)$ is identically zero, thus proving the claim.

We now estimate the number of inequalities which must be satisfied. The number L of vertex and edge labels is $n + (\ell - 2)|E|$, so $L = \mathcal{O}(n^2)$. Then the number M of inequalities is at most $\binom{L}{2} + L^\ell$ which is $\mathcal{O}(n^{2\ell})$.

Order the edges arbitrarily. We choose the values $\mathrm{val}(Z)$ of each variable Z, starting with the vertex labels $x_v = \mathrm{val}(X_v)$ in order, then for each edge e in turn, choosing the $\ell - 3$ free labels $y_e^i = \mathrm{val}(Y_e^i)$ for $i = 1, \ldots, \ell - 3$ (the dependent label y_e^0 will then be

determined by $y_e^0 = (a_1x_v + a_2x_w + b_1y_e^1 + \cdots + b_{\ell-3}y_e^{\ell-3})/b_0)$. We shall choose the value of each label to be an integer multiple of b_0 and we must avoid any value which would make one of the reduced linear combinations equal to zero. When choosing each label, there is some subset of the reduced linear combinations which contain the new label and whose other labels have already been chosen. Each reduced linear combination in this subset forbids one value for the new label (though the forbidden value will be irrelevant if it is not an integer multiple of b_0). Thus each label may be chosen from any set containing at least $M + 1$ positive integer multiples of b_0. We choose the labels so that they are increasing; thus each successive label is chosen from the next available block of $M + 1$ integer multiples of b_0. In particular, none of the labels will be greater than $L(M + 1)b_0$, which is $\mathcal{O}(n^{2\ell+2})$, thus satisfying Condition 5.

Finally, consider any dependent label y_e^0, where $e = (v, w)$ with $v < w$. We need to check that provided not all of the coefficients of $\mathcal{L}$ have the same sign, y_e^0 is a positive integer. Note that in this case, $b_0 > 0$. By definition, $y_e^0 = (a_1x_v + a_2x_w + b_1y_e^1 + \cdots + b_{\ell-3}y_e^{\ell-3})/b_0$. Since all of the other labels are chosen to be integer multiples of b_0, y_e^0 will be an integer. Also, because of the order in which the labels are chosen, we have $x_v < x_w < y_e^1 < \cdots < y_e^{\ell-3}$. Since $a_1 \leq a_2 \leq b_1 \leq \cdots \leq b_{\ell-3}$, and (the negation of) the sum of the negative coefficients is less than the sum of the positive coefficients, it follows that $y_e^0 > 0$, and so $y_e^0 \in \mathbb{N}$. Thus $A \subseteq \mathbb{N}$, as required. □

Note: The exponent $2\ell + 2$ could be reduced to $2\ell - 1$, since for each label the number of linear combinations involving it is $\mathcal{O}(n^{2(\ell-1)})$ and the requirement that the labels be totally ordered can be weakened to the requirement that each edge-set is totally ordered, reducing the factor this contributes from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$.

The following two lemmas relate the size of an $\mathcal{L}$-free subset of the set A constructed above to the size of an independent set in the graph G. These follow very closely the corresponding results in [6], though there a hypergraph construction is used.

Corollary 3.3. *Let $\mathcal{L}$ be an equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero. Let $G = (V, E)$ be a graph and $A = A_V \cup A_E$ the set constructed in Lemma 3.2.*

Then for any $k \in \mathbb{N}$, there is a one-to-one correspondence between independent sets of G of cardinality k and the $\mathcal{L}$-free subsets of A of cardinality $|A_E| + k$ which contain all the elements of A_E.

Proof. The corollary follows immediately from Lemma 3.2. Given an independent set I of G, let $A_I = \{x_v : v \in I\}$, then $A_I \cup A_E$ is an $\mathcal{L}$-free subset of A of size $|I| + |A_E|$; since the vertex labels are distinct, the $\mathcal{L}$-free subsets corresponding to independent sets $I_1 \neq I_2$ are distinct. Also, given any $\mathcal{L}$-free subset $S \cup A_E$ of A of size $|S| + |A_E|$, then $V_S = \{v : x_v \in S\}$ is an independent set in G of size $|S|$, and again we obtain a distinct independent set for each such $\mathcal{L}$-free subset. □

Corollary 3.4. *Let $\mathcal{L}$ be an equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero. Let $G = (V, E)$ be a graph and $A = A_V \cup A_E$ the set constructed in Lemma 3.2.*

Then for any $k \in \mathbb{N}$, G contains an independent set of cardinality k if and only if A contains an $\mathcal{L}$-free subset of cardinality $|A_E| + k$.

Proof. By Corollary 3.3, it suffices to show that, if A contains an $\mathcal{L}$-free subset of cardinality $|A_E| + k$, then in fact A contains such a subset which includes all elements of A_E. To see this, let A_1 be an $\mathcal{L}$-free subset of A of size $|A_E| + k$ which does not contain all elements of A_E; we will show how to construct an $\mathcal{L}$-free subset of equal or greater size which does have this additional property.

Suppose that $y_e^i \in A_E$ but $y_e^i \notin A_1$, and suppose that $e = (v, w)$. Consider the labels x_v, x_w of the endpoints of e. By Lemma 3.2(3) every non-trivial solution to $\mathcal{L}$ involving y_e^i must also involve x_v and x_w. If one of these does not lie in A_1 we add y_e^i to A_1 without creating a solution to $\mathcal{L}$. Otherwise, arbitrarily remove one of x_v and x_w from A_1 and replace it with y_e^i. Repeating this process, we obtain an $\mathcal{L}$-free subset which contains A_E and is at least as large as A_1. □

We can now prove a generalisation of Theorem 2.1. Recall that the problem Independent Set below is NP-complete [5].

Independent Set *Input:* A graph G and integer k. *Question:* Does G have an independent set of size at least k?

Theorem 3.5. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero. Then $\mathcal{L}$-Free Subset is strongly NP-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$.*

Proof. We reduce Independent Set to $\mathcal{L}$-Free Subset. If (G, k) is an instance of Independent Set, we construct (in polynomial time) an instance $(A, |A_E| + k)$ of $\mathcal{L}$-Free Subset, with $A = A_V \cup A_E$, as in Lemma 3.2. Then by Corollary 3.4, G has an independent set of size k if and only if A has a subset of size $|A_E| + k$ with no non-trivial solution to $\mathcal{L}$. The result follows. □

3.1 Sub-equation free sets

In order to prove the later results, it is helpful to have a slightly stronger version of Lemma 3.2.

For a homogeneous equation $\mathcal{L}$, we will call an equation obtained by deleting a non-empty set of the terms of the equation (that is, some variables and their coefficients) a proper sub-equation of $\mathcal{L}$. We will say that a set S is $\mathcal{L}$-proper-sub-equation-free if and only if S does not contain a non-trivial solution to any of the proper sub-equations of $\mathcal{L}$. If S is also $\mathcal{L}$-free, we will say it is $\mathcal{L}$-sub-equation-free.

Making the set A $\mathcal{L}$-proper-sub-equation-free will be useful in Section 5 where we consider the problem ε-$\mathcal{L}$-Free Subset which concerns the size of the largest $\mathcal{L}$-free subset as a proportion of the whole set. To do this we want to extend A by adding new elements while controlling the size of the largest $\mathcal{L}$-free subset, and to achieve this we need to avoid introducing solutions to the equation involving both the original and the new elements. This can be done by scaling the new elements except in the case where these satisfy a sub-equation; in this case we need to ensure that the original elements do not also satisfy

the complementary sub-equation. Note that in the two cases considered in [6], both with three variables, this issue does not really arise because either the sub-equation or its complement has a single variable and so zero is the only solution.

We now show that we can modify Lemma 3.2 to ensure that the set A constructed is $\mathcal{L}$-proper-sub-equation-free. The only change is the addition of Condition 7.

Lemma 3.6. *Consider a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero, and let $\mathcal{L}$ be an equivalent linear equation $a_1x_1 + a_2x_2 + b_1y_1 + \cdots + b_{\ell-3}y_{\ell-3} = b_0y_0$ in standard form. Let $G = (V, E)$ be a graph, and let $n = |V|$. Then we can construct in polynomial time a set $A \subseteq \mathbb{Z}$ with the following properties:*

1. *A is the union of $\ell - 1$ sets $A_V, A_E^0, A_E^1, \ldots, A_E^{\ell-3}$, where $A_V = \{x_v : v \in V\}$ and $A_E^i = \{y_e^i : e \in E\}$ for each $i = 0, 1, \ldots, \ell - 3$;*
2. *$|A| = |V| + (\ell - 2)|E|$;*
3. *for every edge $e = (v, w)$, some permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ is a solution to $\mathcal{L}$; such a solution we call an edge-solution;*
4. *if $(z_1, \ldots, z_\ell)$ is a non-trivial solution to $\mathcal{L}$, then $(z_1, \ldots, z_\ell)$ is a permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ for some edge $e = (v, w)$;*
5. *$\mathrm{maxabs}(A) = \mathcal{O}(|V|^{2\ell+2})$;*
6. *$A \subseteq \mathbb{N}$ unless all the coefficients $c_1, \ldots, c_\ell$ have the same sign.*
7. *A is $\mathcal{L}$-proper-sub-equation-free.*

Proof. The proof is very similar to the proof of Lemma 3.2. In order to eliminate solutions to proper sub-equations, we have to add extra inequalities to be satisfied by the members of A; that is, for any sub-equation $\mathcal{L}'$ with $k < \ell$ variables, and sequence $z = (Z_1, \ldots, Z_k) \in B^k$, we form the corresponding linear combination $\mathrm{lc}(z)$. If $\mathrm{lc}(z)$ is identically zero, this will correspond to a trivial solution to $\mathcal{L}'$ and so can be discarded. Otherwise, we form the reduced linear combination $\mathrm{rlc}(z)$. Then exactly as in Lemma 3.2, we can show that $\mathrm{rlc}(z)$ is not identically zero (in this case there are no edge-solutions since $\mathcal{L}'$ has at most $\ell - 1$ variables). We then choose the set A exactly as before, but now it will also satisfy the extra inequalities which ensure that A is $\mathcal{L}$-proper-sub-equation-free. The total number of inequalities is still $\mathcal{O}(L^\ell)$, where L is the number of variables. □

Using this construction in place of that in Lemma 3.2, we obtain the following NP-completeness result:

Corollary 3.7. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 3$ and the coefficients c_i are all non-zero. Then $\mathcal{L}$-Free Subset is strongly NP-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$. Also the input set A can be restricted to be $\mathcal{L}$-proper-sub-equation-free.*

3.2 Inhomogeneous Equations

We now consider inhomogeneous equations of the form

$$c_1x_1 + \cdots + c_\ell x_\ell = K,$$

where $K \neq 0$. Clearly such an equation $\mathcal{L}$ can only have a solution if K is divisible by $\gcd(c_1, \ldots, c_\ell)$. In this case, we can prove an analogue of Lemma 3.6, provided we restrict attention to tripartite graphs. A different construction, which associates each label with a particular term of the equation, is needed in order to ensure that the dependent edge label will be an integer; in the homogeneous case this was achieved by choosing each label to be a multiple of b_0, but that approach is not sufficient here.

As in the homogeneous case, in order to prove the results of Section 5, we prove a stronger result than Lemma 3.2. In this case we require that there are no solutions in which some variables take values from some fixed $\mathcal{L}$-free set; as before, this is to ensure that when we extend the set A, we do not introduce new solutions to the equation involving both the original and the new elements.

The conditions here are again the same as in Lemma 3.2, with the addition of Condition 7 (and a different bound in Condition 5).

Lemma 3.8. *Let $\mathcal{L}$ be the linear equation $c_1x_1 + \cdots + c_\ell x_\ell = K$, where $\ell \geq 3$ and the coefficients c_i and constant K are all non-zero, with $\gcd(c_1, \ldots, c_\ell)$ a divisor of K. Let $G = (V, E)$ be a tripartite graph, with the vertex set V partitioned into three independent sets V_1, V_2, V_3. Also let S' be a fixed $\mathcal{L}$-free set. Then we can construct in polynomial time a set $A \subseteq \mathbb{Z}$ with the following properties:*

1. *A is the union of $\ell - 1$ sets $A_V, A_E^0, A_E^1, \ldots, A_E^{\ell-3}$, where $A_V = \{x_v : v \in V\}$ and $A_E^i = \{y_e^i : e \in E\}$ for each $i = 0, 1, \ldots, \ell - 3$;*
2. *$|A| = |V| + (\ell - 2)|E|$;*
3. *for every edge $e = (v, w)$, some permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ is a solution to $\mathcal{L}$; such a solution we call an edge-solution;*
4. *if $(z_1, \ldots, z_\ell)$ is a solution to $\mathcal{L}$, then $(z_1, \ldots, z_\ell)$ is a permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ for some edge $e = (v, w)$;*
5. *$\mathrm{maxabs}(A) = \mathcal{O}(|V|^{2\ell})$;*
6. *$A \subseteq \mathbb{N}$ unless all the coefficients of $\mathcal{L}$ have the same sign.*
7. *$A \cup S'$ has no solutions to $\mathcal{L}$ except for those in A.*

Proof. First note that since the equation $\mathcal{L}$ is inhomogeneous, there are no trivial solutions. As in Lemma 3.2, we can construct a set of inequalities which ensure that the vertex and edge labels are all distinct, and there are no solutions to the equation apart from the edge solutions.

Let $g = \gcd(c_1, \ldots, c_\ell)$ so that there exist integers $q_1, \ldots q_\ell$ with

$$c_1q_1 + \cdots + c_\ell q_\ell = g.$$

Also set $K' = K/g$.

In Lemma 3.2 we associated the vertices with the first two coefficients of the equation, however in this case, in order to ensure that the dependent edge labels are integers, we will require that the labels associated with coefficient c_i are congruent to $K'q_i$ modulo $c_1c_2c_3$. This means that each vertex must be associated with a particular coefficient.

As in Lemma 3.2, we will have one vertex label for each vertex. For each edge there will be $\ell - 3$ free edges labels and one dependent edge label. Exactly as before, the set of variables is $B = B_V \cup B_E$, where $B_E = B_E^0 \cup B_E^1 \cup \cdots \cup B_E^{\ell-3}$, $B_V = \{X_v, v \in V\}$ and $B_E^i = \{Y_e^i : e \in E\}$ for each $i = 0, 1, \ldots, \ell - 3$. The variable X_v is associated with the vertex v, and the variables $Y_e^0, \ldots, Y_e^{\ell-3}$ with the edge e.

We now consider the constraints that the variables must satisfy.

First, for each edge $e = (v, w)$, we have a set of variables corresponding to the edge and its endpoints, that is $\{X_v, X_w, Y_e^1, \ldots, Y_e^{\ell-3}, Y_e^0\}$. We will call such a set an edge-set of variables.

For any sequence $z = (Z_1, \ldots, Z_\ell) \in B^\ell$ of variables, we say that z is an edge-sequence if the set $\{Z_1, \ldots, Z_\ell\}$ is an edge-set.

As before, we have the following inequalities:

Type 1: For any distinct pair $z = (Z, Z') \in B^2$, we have the linear combination $\mathrm{lc}(z) = Z - Z' \neq 0$.

Type 2: For any sequence $z = (Z_1, \ldots, Z_\ell) \in (B \cup S')^\ell$ which is not an edge-sequence, we have $\mathrm{lc}(z) = c_1 Z_1 + \cdots + c_\ell Z_\ell - K \neq 0$.

The variables in an edge-set must also satisfy the equation in some order. To ensure this, we associate the three parts V_1, V_2, V_3 of V with coefficients c_1, c_2, c_3. Thus for any edge $e = (v, w)$, the vertices v, w will be in distinct parts, so we associate these with corresponding coefficients, and the dependent edge variable y_e^0 with the third of the coefficients c_1, c_2, c_3. The free edge variables $y_e^1, \ldots, y_e^{\ell-3}$ are associated with $c_4, \ldots, c_\ell$. Thus if v, w are in parts $p(v), p(w)$ respectively and $p(e)$ is the unique element of $\{1, 2, 3\} \setminus \{p(v), p(w)\}$ we have the equation

$$c_{p(v)} X_v + c_{p(w)} X_w + c_{p(e)} Y_e^0 + c_4 Y_e^1 + \cdots + c_\ell Y_e^{\ell-3} = K. \tag{3}$$

Now in each of the linear combinations $\mathrm{lc}(z)$, we substitute for each dependent edge variable Y_e^0 using these equations, and collect terms to obtain a reduced (rational) linear combination $\mathrm{rlc}(z)$ of the vertex variables and free edge variables in the set $B_V \cup B_E^1 \cup \cdots \cup B_E^{\ell-3}$.

Then, as in Lemma 3.2, none of the reduced combinations is identically zero. Hence we can choose values for each of the vertex labels and free edge labels so that all of the inequalities are satisfied, and from these determine the dependent edge labels. We need to check, however, that the dependent edge labels are actually integers. As above, we choose the values so that each label associated with coefficient c_i is congruent to $K' q_i$ modulo $c_1 c_2 c_3$.

So suppose that the dependent edge label y_e^0 of some edge $e = (v, w)$ is associated with coefficient $c_{p(e)}$ (where $p(e) \in \{1, 2, 3\}$). Then from Equation (3) we have

$$\begin{aligned} K - (c_{p(v)} x_v + c_{p(w)} x_w + \sum_{i \geq 4} c_i y_e^{i-3}) &\equiv K - \sum_{i \neq p(e)} c_i K' q_i \pmod{c_1 c_2 c_3} \\ &\equiv K - K'(g - c_{p(e)} q_{p(e)}) \pmod{c_1 c_2 c_3} \\ &\equiv K' c_{p(e)} q_{p(e)} \pmod{c_1 c_2 c_3} \\ &\equiv 0 \pmod{c_{p(e)}}. \end{aligned}$$

Hence $K - (c_{p(v)}x_v + c_{p(w)}x_w + \sum_{i\geq 4} c_i y_e^{i-3})$ is divisible by $c_{p(e)}$ and so y_e^0, which is given by

$$y_e^0 = (K - (c_{p(v)}x_v + c_{p(w)}x_w + \sum_{i\geq 4} c_i y_e^{i-3}))/c_{p(e)},$$

is an integer as required.

Note that the number M of inequalities is $\mathcal{O}(|A|^\ell)$, so the elements of A can be chosen to be at most $\mathcal{O}(|A|^\ell) = \mathcal{O}(|V|^{2\ell})$.

Finally we need to show that if not all of the coefficients $c_1, \ldots, c_\ell$ have the same sign, then we can choose all the labels to be positive.

Note first that for any $T > 0$, we can find a rational solution

$$c_1 m_1 + \ldots + c_\ell m_\ell = K$$

with each $m_i \geq T$. To see this, choose two coefficients $c_I < 0$, $c_J > 0$ and choose each $m_i, i \neq I, J$ so that $m_i \geq T$. Then if $\sum_{i\neq I,J} c_i m_i = N$, we have $m_J = (K - N - c_I m_I)/c_J$ and since $c_I < 0$ it is clear that we can choose m_I, m_J both at least T.

Now if

$$c_1 x_1 + \ldots + c_\ell x_\ell = K$$

is a solution of $\mathcal{L}$ and for some I, we have $|x_i - m_i| \leq r$ for each $i \neq I$, then it follows that

$$|x_I - m_I| \leq (|K| + \sum_{i\neq I} |c_i||x_i - m_i|)/|c_I| \leq (|K| + r\sum_{i\neq I} |c_i|)/|c_I|.$$

Thus by setting $R = (|K| + r\sum_i |c_i|)/(\min_i |c_i|)$, we have $|x_I - m_I| \leq R$. To ensure that all the x_i are positive, it suffices to ensure that $T > R$. We will exploit this by choosing all the vertex and free edge labels to satisfy $|x_i - m_i| \leq r$. Then the dependent edge labels are guaranteed to be positive.

Since there are M inequalities to be satisfied, each label must avoid at most M values. For a label associated with the coefficient c_i, we want to choose it to be positive and congruent to $K'q_i$ modulo $c_1c_2c_3$. Thus if $2r > (M+1)c_1c_2c_3$, then the range $[m_i - r, m_i + r]$ includes at least $M+1$ integers congruent to $K'q_i$ modulo $c_1c_2c_3$, so we can choose the label from among these ensuring that it differs from m_i by at most r. Hence the dependent label will differ from one of m_1, m_2, m_3 by at most R, and so will be positive provided $T > R$. We can achieve this with a value of T which is $\mathcal{O}(M) = \mathcal{O}(|V|^{2\ell})$, hence the elements of A can be positive and polynomially-sized. □

As above we obtain the following:

Corollary 3.9. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = K$, where $\ell \geq 3$ and the coefficients c_i and constant K are all non-zero, with $\gcd(c_1, \ldots, c_\ell)$ a divisor of K. Then $\mathcal{L}$-Free Subset is strongly NP-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$.*

Proof. The proof is similar to that of Theorem 3.5, but the input is restricted to tripartite graphs. We apply Lemma 3.8 (with $S' = \emptyset$) to obtain the required set A. Note that Independent Set is NP-complete for cubic planar graphs [5], so in particular is NP-complete for tripartite graphs. □

4 Approximation

We now consider the complexity of determining the size of the largest solution-free subset of a set. The material in this section is really just a generalisation, to all linear equations in at least three variables, of the corresponding result given by Meeks and Treglown [6], who considered only the equation $a_1x_1 + a_2x_2 = b_0y$ where $a_1, a_2, b_0 \in \mathbb{N}$, so we follow their treatment closely.

We consider the following computational problem:

> 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 $A' \subseteq A$?

If Π is an optimisation problem such as the one above, then for any instance I of Π, we denote by $\mathrm{opt}(I)$ the value of the optimal solution to I (in this example above, the cardinality of the largest solution-free subset). An approximation algorithm $\mathcal{A}$ for Π has *performance ratio* ρ if for any instance I of Π, the value $\mathcal{A}(I)$ given by the algorithm $\mathcal{A}$ satisfies

$$1 \leq \frac{\mathrm{opt}(I)}{\mathcal{A}(I)} \leq \rho.$$

An algorithm $\mathcal{A}$ is a *polynomial-time approximation scheme (PTAS)* for Π if it takes as input an instance I of Π and a real number $\varepsilon > 0$, runs in time polynomial in the size of I (but not necessarily in $1/\varepsilon$) and outputs a value $\mathcal{A}(I)$ such that

$$1 \leq \frac{\mathrm{opt}(I)}{\mathcal{A}(I)} \leq 1 + \varepsilon.$$

We show in this section that for any linear equation $\mathcal{L}$ with at least three variables, there can be no PTAS for the problem MAXIMUM $\mathcal{L}$-FREE SUBSET unless $\mathsf{P} = \mathsf{NP}$.

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 an optimisation problem which is hard for APX does not have PTAS unless $\mathsf{P} = \mathsf{NP}$. 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 $I_2 = f(I_1, \varepsilon)$ of Π_2;
2. if $\varepsilon > 0$ is any constant and y is any solution to I_2 such that $\frac{\mathrm{opt}(I_2)}{|y|} \leq \alpha(\varepsilon)$, then $x = g(I_1, y, \varepsilon)$ is a solution to I_1 such that $\frac{\mathrm{opt}(I_1)}{|x|} \leq 1 + \varepsilon$.

It was shown by Alimonti and Kann [1] that the problem (called MAX IS-3) of determining the maximum size of an independent set in a graph of maximum degree 3 is APX-hard.

Lemma 4.1. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = K$, where $\ell \geq 3$ and the coefficients c_i are all non-zero, and $\gcd(c_1, \ldots, c_\ell)$ is a divisor of K. Then there is a PTAS reduction from* Max IS-3 *to* Maximum $\mathcal{L}$-Free Subset.

Proof. The proof follows closely the proof of Lemma 7 in [6]. 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 $\varepsilon > 0$, outputs the set $A = A_V \cup A_E \subseteq \mathbb{Z}$ described in Lemma 3.2 in the homogeneous case or Lemma 3.8 (with $S' = \emptyset$) in the inhomogeneous case; we know from these lemmas 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. $A_E \subseteq \widetilde{B}$; 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 3.4 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 $\varepsilon > 0$, outputs the set $\{v : x_v \in \widetilde{B} \setminus A_E\}$.

Finally, we define α by $\alpha(\varepsilon) = 1 + \frac{\varepsilon}{6(\ell-2)+1}$. 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 $\mathrm{opt}(A) = \mathrm{opt}(G) + (\ell-2)|E|$. To complete the proof, it suffices to show that whenever B is an $\mathcal{L}$-free subset in A such that $\frac{\mathrm{opt}(A)}{|B|} \leq \alpha(\varepsilon) = 1 + \frac{\varepsilon}{6(\ell-2)+1}$, we have $\frac{\mathrm{opt}(G)}{|I|} \leq 1 + \varepsilon$, where $I = \{v : x_v \in \widetilde{B} \setminus A_E\}$. Observe that

$$
\begin{aligned}
\frac{\mathrm{opt}(A)}{|B|} &\leq 1 + \frac{\varepsilon}{6(\ell-2)+1} \\
\Rightarrow \frac{\mathrm{opt}(A)}{|\widetilde{B}|} &\leq 1 + \frac{\varepsilon}{6(\ell-2)+1} \\
\Rightarrow \frac{(\ell-2)|E| + \mathrm{opt}(G)}{(\ell-2)|E| + |I|} &\leq 1 + \frac{\varepsilon}{6(\ell-2)+1} \\
\Rightarrow \frac{(\ell-2)|E| + \mathrm{opt}(G)}{|I|} &\leq \left(1 + \frac{\varepsilon}{6(\ell-2)+1}\right)\left(\frac{(\ell-2)|E| + |I|}{|I|}\right) \\
\Rightarrow \frac{\mathrm{opt}(G)}{|I|} &\leq \left(1 + \frac{\varepsilon}{6(\ell-2)+1}\right)\frac{(\ell-2)|E|}{|I|} + \left(1 + \frac{\varepsilon}{6(\ell-2)+1}\right) - \frac{(\ell-2)|E|}{|I|} \\
&= \frac{\varepsilon}{6(\ell-2)+1}\frac{(\ell-2)|E|}{|I|} + 1 + \frac{\varepsilon}{6(\ell-2)+1}.
\end{aligned}
$$

Since G has maximum degree 3, we have $|E| \leq \frac{3|V|}{2}$. Also $\widetilde{B}$ is maximal and so I is a maximal independent set, hence I and its neighbours form the whole of V and so $|V| \leq 4|I|$. Thus $|I| \geq \frac{|V|}{4}$ and it follows that $\frac{|E|}{|I|} \leq 6$. We can therefore conclude that

$$
\frac{\mathrm{opt}(A)}{|B|} \leq 1 + \frac{\varepsilon}{6(\ell-2)+1} \Rightarrow \frac{\mathrm{opt}(G)}{|I|} \leq 6(\ell-2)\frac{\varepsilon}{6(\ell-2)+1} + 1 + \frac{\varepsilon}{6(\ell-2)+1} = 1 + \varepsilon,
$$

as required. □

As a corollary, we obtain the following generalisation of Theorem 2.2:

Theorem 4.2. *Let $\mathcal{L}$ be a linear equation with at least three variables. Then* Maximum $\mathcal{L}$-Free Subset *is* APX*-hard.*

5 $\mathcal{L}$-free subsets with a given proportion of elements

In this section we consider the problem of deciding if a set has a solution-free subset containing a given proportion of its elements.

Let $\mathcal{L}$ be the equation

$$c_1x_1 + \cdots + c_\ell x_\ell = K$$

and let $C = \sum_i^\ell c_i$.

If $C \neq 0$ and C divides K, let $S_\mathcal{L} = \{K/C\}$ (otherwise $S_\mathcal{L}$ is empty). Then note that $S_\mathcal{L}$ is an integer set that has no non-empty subset which is $\mathcal{L}$-free; on the other hand, any non-empty subset of $\mathbb{Z} \setminus S_\mathcal{L}$ has a non-empty $\mathcal{L}$-free subset. For this reason it is convenient to exclude this value from the sets of which we seek sum-free subsets. Let $\mathcal{C}(\mathcal{L})$ denote the set of all non-negative reals λ so that every non-empty finite set $Z \subseteq \mathbb{Z} \setminus S_\mathcal{L}$ contains an $\mathcal{L}$-free subset of size strictly greater than $\lambda|Z|$.

Define

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

It follows from the results of Erdős [4] and Eberhard, Green and Manners [2] that for the sum-free equation $\mathcal{L} : x + y = z$, we have $\kappa(\mathcal{L}) = 1/3$.

Given any linear equation $\mathcal{L}$ and a constant rational ε satisfying $0 \leq \varepsilon \leq 1$, we define the following problem.

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

We restrict ε to rational values in order to ensure that the problem is in NP. Note that if $\varepsilon \leq \kappa(\mathcal{L})$, then it follows from the definition of $\kappa(\mathcal{L})$ that every instance of ε-$\mathcal{L}$-Free Subset is a yes-instance, and so the problem is trivially in P in this case.

5.1 Hardness of ε-$\mathcal{L}$-Free Subset

In this section we show that ε-$\mathcal{L}$-Free Subset is strongly NP-complete whenever $\kappa(\mathcal{L}) < \varepsilon < 1$. Given a set $X \subseteq \mathbb{Z}$ and $y \in \mathbb{N}$, we write yX as shorthand for $\{yx : x \in X\}$, and $X + y$ for $\{x + y : x \in X\}$.

Recall that for a homogeneous equation $\mathcal{L}$, we say that a set S is $\mathcal{L}$-sub-equation-free if and only if S does not contain a non-trivial solution to $\mathcal{L}$ or any of its sub-equations.

Lemma 5.1. *Let $\mathcal{L}$ be a translation invariant homogeneous linear equation, and let S be an $\mathcal{L}$-free set of integers. Then for all except a finite number of positive integers α, the set $S + \alpha$ is $\mathcal{L}$-sub-equation-free.*

Proof. Let $\mathcal{L}$ be the equation $c_1x_1 + \cdots + c_\ell x_\ell = 0$, and let $S = \{s_1, \ldots, s_n\}$ be an $\mathcal{L}$-free set. We will say a subsequence $(c_{q_1}, \ldots, c_{q_k})$ of $(c_1, \ldots, c_\ell)$ has non-zero sum if $c_{q_1} + \cdots + c_{q_k} \neq 0$. We choose α to be any positive integer which is not equal to $-(c_{q_1}s_{r_1} + \cdots + c_{q_k}s_{r_k})/(c_{q_1} + \cdots + c_{q_k})$ for any subsequence $c_{q_1}, \ldots, c_{q_k}$ with non-zero sum and any $(s_{r_1}, \ldots, s_{r_k}) \in S^k$.

Then $S + \alpha$ is $\mathcal{L}$-sub-equation-free. For suppose that $\mathcal{L}'$: $c_{q_1}x_{q_1} + \cdots + c_{q_k}x_{q_k} = 0$ is a sub-equation of $\mathcal{L}$, and that $(s_{r_1} + \alpha, \ldots, s_{r_k} + \alpha) \in (S + \alpha)^k$ is a solution to $\mathcal{L}'$. We will show that the solution is trivial. First, $(c_{q_1}, \ldots, c_{q_k})$ cannot have non-zero sum, for then

$$c_{q_1}(s_{r_1} + \alpha) + \cdots + c_{q_k}(s_{r_k} + \alpha) = c_{q_1}s_{r_1} + \cdots + c_{q_k}s_{r_k} + \alpha(c_{q_1} + \cdots + c_{q_k}) \neq 0$$

by the definition of α, a contradiction. So $c_{q_1} + \cdots + c_{q_k} = 0$, which means that

$$c_{q_1}s_{r_1} + \cdots + c_{q_k}s_{r_k} = c_{q_1}(s_{r_1} + \alpha) + \cdots + c_{q_k}(s_{r_k} + \alpha) = 0.$$

Since $\mathcal{L}$ is translation invariant, the sum of the remaining coefficients, $\sum_{i \notin \{q_1, \ldots, q_k\}} c_i$, is also zero. Thus if s is any element of S, we have

$$c_{q_1}s_{r_1} + \cdots + c_{q_k}s_{r_k} + \sum_{i \notin \{q_1, \ldots, q_k\}} c_i s = 0.$$

Since S is $\mathcal{L}$-free, this solution of $\mathcal{L}$ must be trivial, which means that the total coefficient of each s_i in $c_{q_1}s_{r_1} + \cdots + c_{q_k}s_{r_k}$ must be zero. Hence the total coefficient of each $s_i + \alpha$ in $c_{q_1}(s_{r_1} + \alpha) + \cdots + c_{q_k}(s_{r_k} + \alpha)$ is zero, so the solution $(s_{r_1} + \alpha, \ldots, s_{r_k} + \alpha)$ of $\mathcal{L}'$ is trivial, as required. □

We now show how to extend an $\mathcal{L}$-free set (provided certain conditions are met). This will be useful to create sets whose largest $\mathcal{L}$-free subset is of some desired size relative to the whole set.

We say that an injective function $f : \mathbb{Z} \to \mathbb{Z}$ is solution-preserving for $\mathcal{L}$ if the following holds: $(f(x_1), \ldots, f(x_\ell))$ is a solution to $\mathcal{L}$ if and only if $(x_1, \ldots, x_\ell)$ is a solution to $\mathcal{L}$.

Lemma 5.2. *Let $\mathcal{L}$ be the linear equation $c_1x_1 + \ldots + c_\ell x_\ell = K$ with $\ell \geq 3$. Let S be a fixed non-empty subset of $\mathbb{Z} \setminus S_{\mathcal{L}}$, $S' \subseteq S$ a fixed non-empty $\mathcal{L}$-free subset of S and A be a subset of $\mathbb{Z} \setminus S_{\mathcal{L}}$ with the following properties.*

- *If $K = 0$ and $\mathcal{L}$ is translation invariant, then S' is $\mathcal{L}$-sub-equation-free and $0 \notin S$.*
- *If $K = 0$, then A is $\mathcal{L}$-proper-sub-equation-free.*
- *If $K \neq 0$, then $A \cup S'$ has no solutions to $\mathcal{L}$ except for those in A.*

Let r and t be non-negative integers. Then we can choose solution-preserving functions $f_1, \ldots, f_r$ and a set $T = \{u_1, \ldots, u_t\}$ comprising t distinct strictly positive integers such that the sets A, $f_i(S)$ for each $i = 1, \ldots, r$, and T are all mutually disjoint and such that

$$A \cup \bigcup_{i=1}^{r} f_i(S') \cup T$$

is a subset of $\mathbb{Z} \setminus S_{\mathcal{L}}$ *and has no non-trivial solutions to* $\mathcal{L}$ *except those in* A*. In particular, if* $A' \subseteq A$ *is* $\mathcal{L}$*-free, then* $A' \cup \bigcup_{i=1}^{r} f_i(S') \cup T$ *is* $\mathcal{L}$*-free. Also the functions* $f_1, \ldots, f_r$ *are polynomially bounded and the elements of* T *need be no larger than* $\mathcal{O}(|A| + r + t)^{\ell}$*.*

Proof. We will use a similar technique to that used in Lemma 3.2, that is, we will use a variable corresponding to each integer that needs to be chosen, and obtain a set of inequalities in these variables. Then choosing values of the variables to satisfy all of the inequalities will give the result.

Let $S = \{s_1, \ldots, s_k\}$, and assume that $S' = \{s_1, \ldots, s_{k'}\}$. There are several cases to consider.

Case 1. The equation $\mathcal{L}$ is homogeneous, that is, $K = 0$.

Note that by assumption, this means that A is $\mathcal{L}$-proper-sub-equation-free. In this case we will use the functions $f_i(s) = d_i s$, where $d_1, \ldots, d_r$ are integers to be chosen.

We will take a set $D = \{D_1, \ldots, D_r\}$ of variables, with variable D_i corresponding to integer d_i, and a set $U = \{U_1, \ldots, U_t\}$ of variables with U_i corresponding to element u_i. We will also write s_{ij} for $d_i s_j$ and take a variable S_{ij} corresponding to s_{ij}; the variable S_{ij} of course depends on D_i since $S_{ij} = D_i s_j$. Write $\Sigma = \{S_{ij} : 1 \leq i \leq r, 1 \leq j \leq k\}$ and $\Sigma' = \{S_{ij} : 1 \leq i \leq r, 1 \leq j \leq k'\}$.

Now for any distinct pair $z = (Z, Z') \in (A \cup \Sigma \cup U)^2 \setminus A^2$, we form the combination $\mathrm{lc}(z) = Z - Z'$. Note that we require that all the new elements are distinct, so Z and Z' range over $A \cup \Sigma \cup U$ and not merely $A \cup \Sigma' \cup U$.

Also for any sequence $z = (Z_1, \ldots, Z_\ell) \in (A \cup \Sigma' \cup U)^\ell \setminus A^\ell$, we form the combination $\mathrm{lc}(z) = c_1 Z_1 + \ldots + c_\ell Z_\ell$. Note that we only include sequences with at least one variable from $\Sigma' \cup U$, and in this case we do not consider variables from $\Sigma \setminus \Sigma'$.

For each sequence z, we substitute $S_{ij} = D_i s_j$ for any variables in Σ to obtain the reduced combination $\mathrm{rlc}(z)$.

We will say that $\mathrm{rlc}(z)$ is identically zero if both the constant part (which is a linear combination of elements from A) and the total coefficient of each variable, are zero. We will say that $\mathrm{lc}(z)$ is trivial if the total coefficient of each element of $A \cup \Sigma \cup U$ in $\mathrm{lc}(z)$ is zero. Note this is stronger than saying that $\mathrm{lc}(z)$ is identically zero, because we require that the total coefficient of each element of A is zero, not merely that the total constant term is zero. The combination $\mathrm{lc}(z)$ is trivial if and only if z corresponds to a trivial solution of $\mathcal{L}$, that is, the total coefficient of each value is zero.

Claim. If $\mathrm{rlc}(z)$ is identically zero, then $\mathrm{lc}(z)$ is trivial.

Equivalently, either $\mathrm{rlc}(z)$ is not identically zero, or $\mathrm{lc}(z)$ is trivial.

For distinct pairs $z = (Z, Z') \in (A \cup \Sigma \cup U)^2 \setminus A^2$ this is clear, since the only case where the terms in $\mathrm{lc}(z)$ could cancel after substitution is if $Z = S_{ij}$ and $Z' = S_{i'j'}$. But then $Z - Z' = S_{ij} - S_{i'j'}$ which reduces to $\mathrm{rlc}(z) = s_j D_i - s_{j'} D_{i'}$. If this is identically zero then we must have $s_j = s_{j'}$ so $\mathrm{rlc}(z) = s_j(D_i - D_{i'})$. Since $0 \notin S$ (if $\mathcal{L}$ is translation invariant, this is by assumption, otherwise because $S_{\mathcal{L}} = \{0\}$), this means that $D_i = D_{i'}$ also, contradicting the assumption that Z, Z' are distinct.

Now consider any sequence $z = (Z_1, \ldots, Z_\ell) \in (A \cup \Sigma' \cup U)^\ell \setminus A^\ell$, so that $\mathrm{lc}(z) = c_1 Z_1 + \ldots + c_\ell Z_\ell$. Note that $\mathrm{lc}(z)$ is the sum of three parts, an A-part, which is an integer

formed from a linear combination of elements of A, a Σ'-part which is a linear combination of variables in Σ', and a U-part which is a linear combination of variables in U. Similarly $\mathrm{rlc}(z)$ is the sum of an A-part, a D-part and a U-part. Note that since the substitution only affects the variables in Σ', the A-parts of $\mathrm{lc}(z)$ and $\mathrm{rlc}(z)$ are equal, and similarly for the U-parts. For any D-variable D_i, the coefficient of D_i in $\mathrm{rlc}(z)$ is a linear combination of the form $c_{q_1}s_{r_1} + \ldots + c_{q_k}s_{r_k}$, where each $s_{r_j} \in S'$.

Now if $\mathrm{rlc}(z)$ is identically zero, its A-part must be zero. This gives a solution to a sub-equation of $\mathcal{L}$, which must be trivial since A is $\mathcal{L}$-proper-sub-equation-free. Thus for each element of the sequence z which comes from the set A, its total coefficient in $\mathrm{rlc}(z)$, and therefore also in $\mathrm{lc}(z)$, must be zero.

Also if $\mathrm{rlc}(z)$ is identically zero, its U-part must be identically zero, hence the same is true for $\mathrm{lc}(z)$, that is, every U-variable in z has total coefficient zero in $\mathrm{lc}(z)$. Now there are two cases, depending on whether or not $\mathcal{L}$ is translation invariant.

Subcase 1.1. $\mathcal{L}$ is not translation invariant.

In this case $\mathrm{lc}(z)$ cannot be trivial.

We must show that $\mathrm{rlc}(z)$ is not identically zero. So suppose that it is. Pick a fixed $s \in S'$, and in $\mathrm{rlc}(z)$, replace each element of A with the value s, and set each D-variable (temporarily) to 1 and each U-variable (temporarily) to s. Since $\mathrm{rlc}(z)$ is identically zero, then from above the total coefficients of its A- and U-parts are both zero, and this results in a solution to $\mathcal{L}$ in S', contradicting the assumption that S' is $\mathcal{L}$-free.

Subcase 1.2. $\mathcal{L}$ is translation invariant.

In this case, by assumption, S' is $\mathcal{L}$-sub-equation-free.

Suppose that $\mathrm{rlc}(z)$ is identically zero. Then we must show that $\mathrm{lc}(z)$ is trivial. From above, the total coefficient in $\mathrm{lc}(z)$ of each element of A and each U-variable is zero, so to show that $\mathrm{lc}(z)$ is trivial, we need to show that the total coefficient in $\mathrm{lc}(z)$ of each Σ'-variable is also zero. We do this as follows. Since $\mathrm{rlc}(z)$ is identically zero, the total coefficient of each D-variable in $\mathrm{rlc}(z)$ is zero.

Consider the coefficient of some D-variable D_i. This coefficient is a linear combination of the form $c_{q_1}s_{r_1} + \ldots + c_{q_k}s_{r_k}$, where each $s_{r_j} \in S'$. As the coefficient of D_i is 0, this gives a solution to a sub-equation of $\mathcal{L}$, so since S' is $\mathcal{L}$-sub-equation-free, this solution is trivial and so the total coefficent of each s_j in $c_{q_1}s_{r_1} + \ldots + c_{q_k}s_{r_k}$ is zero. But this is just the coefficient of $s_jD_i = S_{ij}$ in $\mathrm{lc}(z)$, hence $\mathrm{lc}(z)$ is trivial, as required.

Hence whenever $\mathrm{lc}(z)$ is not trivial, $\mathrm{rlc}(z)$ is not identically zero. This completes the proof of the claim.

We form a set of inequalities by including the inequality $\mathrm{rlc}(z) \neq 0$ whenever $\mathrm{lc}(z)$ is non-trivial.

Now as in Lemma 3.2, we choose values $d_i = \mathrm{val}(D_i)$ and $u_i = \mathrm{val}(U_i)$ for each of the variables in $D \cup U$, from which those in Σ follow by setting $s_{ij} = d_is_j$. We can choose the values of the D-variables and U-variables to ensure that the value $\mathrm{val}(\mathrm{rlc}(z))$ of $\mathrm{rlc}(z)$ is non-zero. But since clearly $\mathrm{val}(\mathrm{rlc}(z) = \mathrm{val}(\mathrm{lc}(z))$, this ensures that there are no non-trivial solutions to $\mathcal{L}$ in $A \cup \bigcup_{i=1}^{r} d_iS' \cup T$ except those already present in A. In particular, if $A' \subseteq A$ is $\mathcal{L}$-free, then so is $A' \cup \bigcup_{i=1}^{r} d_iS' \cup T$.

Case 2. The equation $\mathcal{L}$ is inhomogeneous, that is, $K \neq 0$.

This case is simpler because there are no trivial solutions. Note that, by assumption, $A \cup S'$ has no solutions to $\mathcal{L}$ except for those in A.

As before we choose functions $f_1, \ldots, f_r$. These depend on r integers $d_1, \ldots, d_r$ which we will choose as above using r variables $D_1, \ldots, D_r$.

Let $C = c_1 + \cdots + c_\ell$ be the sum of the coefficients. The solution-preserving functions functions f_i depend on whether C is zero or not. If $C = 0$, let f_i be the function defined by $f_i(s) = s + d_i$. If $C \neq 0$, define f_i by $f_i(s) = (d_i C + 1)s - d_i K$. Now as above set $s_{ij} = f_i(s_j)$, and let S_{ij} be a variable corresponding to s_{ij}, so that $S_{ij} = s_j + D_i$ if $C = 0$ and $S_{ij} = (D_i C + 1)s_j - D_i K = (s_j C - K)D_i + s_j$ otherwise. As above $\Sigma = \{S_{ij} : 1 \leq i \leq r, 1 \leq j \leq k\}$ and $\Sigma' = \{S_{ij} : 1 \leq i \leq r, 1 \leq j \leq k'\}$.

Now as before, for any distinct pair $z = (Z, Z') \in (A \cup \Sigma \cup U)^2 \setminus A^2$, we form the combination $\mathrm{lc}(z) = Z - Z'$, and for any sequence $z = (Z_1, \ldots, Z_\ell) \in (A \cup \Sigma' \cup U)^\ell \setminus A^\ell$, we form the combination $\mathrm{lc}(z) = c_1 Z_1 + \ldots + c_\ell Z_\ell - K$.

As before we substitute for each S_{ij} to obtain the reduced combinations $\mathrm{rlc}(z)$. Note that the constant part of $\mathrm{rlc}(z)$ is now a linear combination of elements from A and S.

Claim. $\mathrm{rlc}(z)$ is not identically zero.

As before, for a distinct pair $z = (Z, Z')$, a problem could only arise if $Z = S_{ij}$ and $Z' = S_{i'j'}$. Then $\mathrm{lc}(z) = Z - Z'$. If $C = 0$ then $\mathrm{rlc}(z) = s_j + D_i - (s_{j'} + D_{i'})$, which can only be identically zero if $s_j = s_{j'}$ and $D_i = D_{i'}$, so Z and Z' are the same variable. If $C \neq 0$ then $\mathrm{rlc}(z) = (s_j C - K)D_i + s_j - ((s_{j'} C - K)D_{i'} + s_{j'})$. If this is identically zero, we must have $s_j = s_{j'}$, and then $\mathrm{rlc}(z) = (s_j C - K)(D_i - D_{i'})$. Since $K/C \notin S$, this can only be identically zero if $D_i = D_{i'}$ but then Z and Z' are the same variable, giving a contradiction.

So suppose that z is a sequence $(Z_1, \ldots, Z_\ell) \in (A \cup \Sigma' \cup U)^\ell \setminus A^\ell$. Note that some Z_i is in $\Sigma' \cup U$. We substitute for each Σ'-variable to obtain $\mathrm{rlc}(z)$. Recall that these substitions are either $S_{ij} = s_j + D_i$ or $S_{ij} = (s_j C - K)D_i + s_j$. Then $\mathrm{rlc}(z)$ has a constant part, a D-part and a U-part. The constant part is of the form $c_{q_1} x_{q_1} + \ldots + c_{q_k} x_{q_k} - K$, where each x_{q_j} is in $A \cup S'$. Then observe that the U-part is $\sum_{i \notin \{q_1, \ldots, q_k\}} c_i Z_i$, where each $Z_i \in U$. Also either some x_{q_j} is in S', or the U-part has at least one term. If $\mathrm{rlc}(z)$ is identically zero, then the constant part must be zero and the U-part identically zero. But then setting (temporarily) each U-variable equal to some fixed element of S' and putting together the constant part and the U-part gives a solution to $\mathcal{L}$ in $A \cup S'$ involving at least one term in S', so this solution is not in A, a contradiction.

Thus $\mathrm{rlc}(z)$ is not identically zero, which proves the claim. So in the inhomogeneous case we can choose values of the D- and U-variables to ensure that every $\mathrm{lc}(z)$ is non-zero, and then there are no additional solutions to $\mathcal{L}$ in $A \cup \bigcup_{i=1}^{r} f_i(S') \cup T$.

Finally, the number of elements in $A \cup D \cup U$ is $|A| + r + t$, hence the number of inequalities is $\mathcal{O}(|A| + r + t)^\ell$. Hence each value chosen has to avoid up to this many values, so we can choose each element of $D \cup U$ to be no larger than $\mathcal{O}(|A| + r + t)^\ell$, with the elements of U being positive.

□

We can now state and prove the NP-completeness result for ε-$\mathcal{L}$-Free Subset. Again this result is a generalisation of one given by Meeks and Treglown [6] and the proof closely follows theirs.

Theorem 5.3. *Let $\mathcal{L}$ be a linear equation with at least three variables, and let ε be a rational number satisfying $0 \leq \varepsilon \leq 1$. Then ε-$\mathcal{L}$-Free Subset is strongly NP-complete if $\kappa(\mathcal{L}) < \varepsilon < 1$, otherwise it is in P.*

Proof. Let $\mathcal{L}$ be the linear equation $c_1x_1 + \ldots + c_\ell x_\ell = K$ with $\ell \geq 3$.

As noted earlier, if $\varepsilon \leq \kappa(\mathcal{L})$, then the problem is trivially in P. Also if $\varepsilon = 1$, then we simply need to check if the set A is $\mathcal{L}$-free, which can also be done in polynomial time.

So assume that $\kappa(\mathcal{L}) < \varepsilon < 1$. By definition of $\kappa(\mathcal{L})$, there is a non-empty set $S \subseteq \mathbb{Z} \setminus S_{\mathcal{L}}$ such that the largest $\mathcal{L}$-free subset S' of S has size precisely $\varepsilon'|S|$ where $\kappa(\mathcal{L}) \leq \varepsilon' < \varepsilon$. Note that S' cannot be empty, because any non-empty subset of $\mathbb{Z} \setminus S_{\mathcal{L}}$ has a non-empty $\mathcal{L}$-free subset. Note also that since ε is a constant here, S and S' can be considered fixed sets and $|S|$ a constant. If $\mathcal{L}$ is homogeneous and translation invariant, then by Lemma 5.1 we can find an integer α such that $S'+\alpha$ is $\mathcal{L}$-sub-equation-free and $0 \notin S+\alpha$. Clearly the largest $\mathcal{L}$-free subset of $S+\alpha$ has size $\varepsilon'|S+\alpha|$. Thus in the case when $\mathcal{L}$ is homogeneous and translation invariant we can assume (by replacing S by $S+\alpha$ and S' by $S'+\alpha$) that S' is $\mathcal{L}$-sub-equation-free and $0 \notin S$.

It is shown in [6] that ε-$\mathcal{L}$-Free Subset belongs to NP (and this is unaffected by the slight change in the defnition). To show that the problem is NP-complete, we describe a reduction from $\mathcal{L}$-Free Subset, shown to be NP-complete in Theorems 3.5 and Corollary 3.9.

Suppose that (A, k) is such an instance of $\mathcal{L}$-Free Subset where $A \subseteq \mathbb{N}$. By Corollary 3.7 and Lemma 3.8, we can assume that the elements of A are bounded by a polynomial in $|A|$, and that (i) if $K = 0$ then A is $\mathcal{L}$-proper-sub-equation-free and (ii) if $K \neq 0$ then $A \cup S'$ has no solution to $\mathcal{L}$ except for those in A. It also follows from Corollary 3.7 and Condition 4 of Lemma 3.8 that we can assume that $A \subseteq \mathbb{Z} \setminus S_{\mathcal{L}}$. We will define a set $B \subseteq \mathbb{Z} \setminus S_{\mathcal{L}}$ 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. Note that we may assume that $k \leq |A|$, since otherwise the instance (A, k) is a "no"-instance and we can take B to be the set S, which is a "no"-instance of ε-$\mathcal{L}$-Free Subset.

Let $a = |A|$, and set

$$r = \max\left(0, \left\lceil \frac{k - \varepsilon a}{(\varepsilon - \varepsilon')|S|} \right\rceil\right),$$

and note that $r = \mathcal{O}(|A|)$. Let $k^* = k + r|S'| = k + r\varepsilon'|S|$ and $a^* = a + r|S|$. Then by definition of r, $r\varepsilon|S| - r\varepsilon'|S| \geq k - \varepsilon a$, so that $k + r\varepsilon'|S| \leq \varepsilon a + r\varepsilon|S| = \varepsilon(a + r|S|)$, or $k^* \leq \varepsilon a^*$.

Now set

$$t = \left\lceil \frac{\varepsilon a^* - k^*}{1 - \varepsilon} \right\rceil,$$

and again note that $t = \mathcal{O}(|A|)$. Then

$$t \geq \frac{\varepsilon a^* - k^*}{1 - \varepsilon} \geq t - 1,$$

so

$$t(1-\varepsilon) \geq \varepsilon a^* - k^* \geq (t-1)(1-\varepsilon),$$

and so

$$k^* + t \geq \varepsilon(a^* + t) \geq k^* + t - (1-\varepsilon).$$

Hence $\lceil \varepsilon(a^* + t) \rceil = k^* + t$, or $\lceil \varepsilon(|A| + r|S| + t) \rceil = k + r|S'| + t$.

By Lemma 5.2, we can construct, in time polynomial in size(A), a set $B = A \cup \bigcup_{i=1}^{r} f_i(S) \cup T$, where the functions f_i are solution-preserving, such that $|B| = |A| + r|S| + t$ and such that $A \cup \bigcup_{i=1}^{r} f_i(S') \cup T$ has no non-trivial solutions to $\mathcal{L}$ apart from those in A.

Then we claim 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.

To show this, first suppose that B has an $\mathcal{L}$-free subset B' of size at least $\varepsilon|B|$. Then B' has size at least $\lceil \varepsilon|B| \rceil = \lceil \varepsilon(|A| + r|S| + t) \rceil = k + r|S'| + t$. Clearly the largest $\mathcal{L}$-free set in $f_i(S)$ is of size $|S'|$. Hence at most $r|S'| + t$ of the elements of B' can be in $\bigcup_{i=1}^{r} f_i(S) \cup T$ so at least k are in A, as required.

On the other hand, suppose that A has an $\mathcal{L}$-free subset A' of size k. Then by Lemma 5.2, the set $A' \cup \bigcup_{i=1}^{r} f_i(S') \cup T$ is $\mathcal{L}$-free, and this set has size $k + r|S'| + t = \lceil \varepsilon(|A| + r|S| + t) \rceil = \lceil \varepsilon|B| \rceil$.

Hence B is a yes-instance for ε-$\mathcal{L}$-Free Subset if and only if (A, k) is a yes-instance for $\mathcal{L}$-Free Subset. □

6 Counting the number of $\mathcal{L}$-free subsets

We move on to consider the counting version of $\mathcal{L}$-Free Subset.

> #$\mathcal{L}$-Free Subset
> *Input:* A finite set $A \subseteq \mathbb{Z}$.
> *Output:* The number of $\mathcal{L}$-free subsets of A.

We first consider the homogeneous case. Here we use two different constructions, with one construction covering equations with three variables and the other equations with four or more variables.

Each construction requires another modified version of Lemma 3.2.

Lemma 6.1. *Consider a linear equation $c_1x_1 + c_2x_2 + c_3x_3 = 0$, where the coefficients c_i are all non-zero, and let $\mathcal{L}$ be an equivalent linear equation $a_1x_1 + a_2x_2 = b_0y_0$ in standard form. Let $G = (V, E)$ be a graph and let $r \in \mathbb{N}$. Then we can construct in polynomial time, a set $A \subseteq \mathbb{Z}$ with the following properties.*

1. *A is the union of three sets A_V, A_E and U, where $A_V = \{x_v : v \in V\}$, $A_E = \{y_e : e \in E\}$ and $U = \{u_{v,e,i} : v \in V, e = vw \in E, 1 \leq i \leq r\}$;*
2. *$|A| = |V| + (1 + 2r)|E|$;*
3. *for every edge $e = vw$, some permutation of (x_v, x_w, y_e) is a solution to $\mathcal{L}$; such a solution we call an edge solution;*

4. *for every edge $e = vw$ and every integer i with $1 \leq i \leq r$, some permutation of $(u_{v,e,i}, u_{w,e,i}, y_e)$ is a solution to $\mathcal{L}$;*

5. *if (z_1, z_2, z_3) is a non-trivial solution to $\mathcal{L}$, then (z_1, z_2, z_3) is a permutation of (x_v, x_w, y_e) for some edge $e = vw$ or of $(u_{v,e,i}, u_{w,e,i}, y_e)$ for some edge $e = vw$ and integer i with $1 \leq i \leq r$;*

6. $\text{maxabs}(A) = \mathcal{O}(|V|^{12} r^2)$*;*

7. $A \subseteq \mathbb{N}$ *unless all the coefficients c_1, c_2, c_3 have the same sign.*

Proof. Apply Lemma 3.2 to produce a set A' with $A'_V = \{x'_v : v \in V\}$ and $A'_E = \{y_e : e \in E\}$. We will modify A'_V and A'_E by multiplying each element of both sets by a large positive integer N (to be chosen later) to give sets A_V and A_E, and then extend $A_V \cup A_E$ by adding, for each edge (v, w), r new elements close to each of the vertex labels $x_v = Nx'_v$ and $x_w = Nx'_w$. Intuitively, these extra elements can be thought of as 'perturbations' of the elements of A_V. We shall see that the only solutions to $\mathcal{L}$ in A come from 'perturbing' a solution to $\mathcal{L}$ comprising only elements of $A_V \cup A_E$.

For an integer M, let $S_M = \{a_2 M, -a_1 M, 0\}$. We first choose a collection of pairwise distinct strictly positive integers $N_{e,i}$, one for each edge e and integer i with $1 \leq i \leq r$, so that for any solution (z_1, z_2, z_3) to $\mathcal{L}$ from the set

$$S = \bigcup_{e \in E, 1 \leq i \leq r} S_{N_{e,i}},$$

we have $\{z_1, z_2, z_3\} \subseteq S_{N_{e,i}}$ for some e and i. Equivalently for all pairwise distinct (e, i), (e', i'), (e'', i''), there is no solution to $\mathcal{L}$ in S including multiples of at least two of $N_{e,i}$, $N_{e',i'}$ and $N_{e'',i''}$. This condition can by enforced by requiring that for any triple (z_1, z_2, z_3) of elements from S including multiples of two different integers $N_{e,i}$, we have $a_1 z_1 + a_2 z_2 \neq b_0 z_3$. For each e and i, there are $\mathcal{O}(|E|^2 r^2)$ such linear inequalities including $N_{e,i}$.

We may choose the integers $N_{e,i}$ greedily in any order. As well as the $\mathcal{O}(|E|^2 r^2)$ values forbidden by the linear inequalities described above, when $N_{e,i}$ is chosen there are an additional $\mathcal{O}(|E|r)$ linear inequalities it must satisfy in order to ensure that it is distinct from all previously chosen values. Thus the integers $N_{e,i}$ may be chosen so that $N' = \max_{e \in E,\ 1 \leq i \leq r} N_{e,i}$ satisfies $N' = \mathcal{O}(|E|^2 r^2)$.

Now let $c = \max\{|c_1|, |c_2|, |c_3|\}$ and $N = 3c^2 N' + 1$. For each $v \in V$, let $x_v = Nx'_v$, and for each $e \in E$ let $y_e = Ny'_e$. Form A_V and A_E by setting $A_V = \{x_v : v \in V\}$ and $A_E = \{y_e : e \in E\}$. Solutions to $\mathcal{L}$ in $A_V \cup A_E$ are in one-to-one correspondence with solutions to $\mathcal{L}$ in $A'_V \cup A'_E$.

We now construct the set U. Start with U being empty, and for each edge of G, add $2r$ elements to it. Suppose that edge e of G joins v and w with v coming before w in the vertex ordering of Lemma 3.2. For each $i = 1, \ldots, r$, add the elements $u_{v,e,i} = x_v + N_{e,i} a_2$ and $u_{w,e,i} = x_w - N_{e,i} a_1$ to U. Finally set $A = A_V \cup A_E \cup U$.

Notice that all the conditions on A in the statement of the lemma are satisfied except possibly Condition 5.

Recall our earlier assertion that solutions to $\mathcal{L}$ in A are 'perturbations' of solutions to $\mathcal{L}$ in $A_V \cup A_E$. We now make this idea more precise, by defining an element $\bar{z}$ of $A_V \cup A_E$

corresponding to each element z of A and showing that if (z_1, z_2, z_3) is a solution to $\mathcal{L}$ in A, then $(\bar{z}_1, \bar{z}_2, \bar{z}_3)$ is also a solution to $\mathcal{L}$.

Given a 3-tuple (z_1, z_2, z_3), let

$$\mathcal{L}(z_1, z_2, z_3) = a_1 z_1 + a_2 z_2 - b_0 z_3.$$

If $z \in A$, then define $\bar{z} = z$ if $z \notin U$ and otherwise if $z = x_v \pm N_{e,i} a_k$ then define $\bar{z} = x_v$. For any 3-tuple (z_1, z_2, z_3) of elements of A,

$$|\mathcal{L}(z_1, z_2, z_3) - \mathcal{L}(\bar{z}_1, \bar{z}_2, \bar{z}_3)| \leq 3c^2 N' < N. \tag{4}$$

Suppose that (z_1, z_2, z_3) is a solution to $\mathcal{L}$ in A. Then $\mathcal{L}(z_1, z_2, z_3) = 0$. Furthermore, each of $\bar{z}_1, \bar{z}_2, \bar{z}_3$ is divisible by N, so $\mathcal{L}(\bar{z}_1, \bar{z}_2, \bar{z}_3)$ is divisible by N. Combining this observation with (4), we deduce that $\mathcal{L}(\bar{z}_1, \bar{z}_2, \bar{z}_3) = 0$. So $(\bar{z}_1, \bar{z}_2, \bar{z}_3)$ is a solution to $\mathcal{L}$ in $A_V \cup A_E$, and we observed earlier that by dividing each element by N we obtain from it a solution to $\mathcal{L}$ in $A'_V \cup A'_E$. Hence either $\{\bar{z}_1, \bar{z}_2, \bar{z}_3\} = \{x_v, x_w, y_e\}$ for some edge $e = vw$ or $\mathcal{L}$ is translation invariant and $(\bar{z}_1, \bar{z}_2, \bar{z}_3) = (z, z, z)$ for some $z \in A_V \cup A_E$.

If $\{\bar{z}_1, \bar{z}_2, \bar{z}_3\} = \{x_v, x_w, y_e\}$, then by permuting the indices if necessary, we may assume that $\bar{z}_1 = x_v$, $\bar{z}_2 = x_w$ and $\bar{z}_3 = y_e$ with v coming before w in the vertex ordering of Lemma 3.2. Then the choice of the integers $N_{e,i}$ ensures that there exists an edge e and integer i such that $z_1 \in \{x_v, x_v + a_2 N_{e,i}\}$, $z_2 \in \{x_w, x_w - a_1 N_{e,i}\}$ and $z_3 = y_e$. But no permutation of $(x_v, x_w - a_1 N_{e,i}, y_e)$ or of $(x_v + a_2 N_{e,i}, x_w, y_e)$ can be a solution of $\mathcal{L}$ because for each permutation σ of $\{1, 2, 3\}$, we have either $\mathcal{L}(\bar{z}_{\sigma(1)}, \bar{z}_{\sigma(2)}, \bar{z}_{\sigma(3)}) = 0$ or $|\mathcal{L}(\bar{z}_{\sigma(1)}, \bar{z}_{\sigma(2)}, \bar{z}_{\sigma(3)}| \geq N$. Thus $(z_1, z_2, z_3) = (x_v, x_w, y_e)$ or $(z_1, z_2, z_3) = (x_v + a_2 N_{e,i}, x_w - a_1 N_{e,i}, y_e)$.

If $(\bar{z}_1, \bar{z}_2, \bar{z}_3) = (z, z, z)$, then either $z = y_e$ for some edge e or $z = x_v$ for some vertex v. If $z = y_e$, then $z_1 = z_2 = z_3 = y_e$ and we have a trivial solution. For any vertex v, edge e and integer i, the set A contains at most one of $x_v + a_2 N_{v,e,i}$ and $x_v - a_1 N_{v,e,i}$. So either $\{z_1, z_2, z_3\} \subseteq \{x_v, x_v + a_2 N_{v,e,i}\}$ or $\{z_1, z_2, z_3\} \subseteq \{x_v, x_v - a_1 N_{v,e,i}\}$ for some edge e and integer i. As $\mathcal{L}$ is translation invariant we have $c_j + c_k \neq 0$ for all j and k. Thus in either case $z_1 = z_2 = z_3$ and we have a trivial solution. Thus Condition 5 holds. □

For the case when $\ell \geq 4$ we have a different construction.

Lemma 6.2. *Consider a linear equation $c_1 x_1 + \cdots + c_\ell x_\ell = 0$, where $\ell \geq 4$ and the coefficients c_i are all non-zero, and let $\mathcal{L}$ be an equivalent linear equation $a_1 x_1 + a_2 x_2 + b_1 y_1 + \cdots + b_{\ell-3} y_{\ell-3} = b_0 y_0$ in standard form. Let $G = (V, E)$ be a graph and let $r \in \mathbb{N}$. Then we can construct in polynomial time a set $A \subseteq \mathbb{Z}$ with the following properties:*

1. *A is the union of sets A_V, A_E, where $A_V = \{x_v : v \in V\}$ and $A_E = \{y^i_{e,j} : e \in E, i = 0, 1, \ldots, \ell - 3, j = 1, \ldots, r\}$; thus for each edge e we have r sets of values;*
2. *$|A| = |V| + (\ell - 2)r|E|$;*
3. *for every edge $e = (v, w)$ and each $j = 1, \ldots, r$, some permutation of $(x_v, x_w, y^1_{e,j}, \ldots, y^{\ell-3}_{e,j}, y^0_{e,j})$ is a solution to $\mathcal{L}$; such a solution we call an edge-solution;*
4. *if $(z_1, \ldots, z_\ell)$ is a non-trivial solution to $\mathcal{L}$, then $(z_1, \ldots, z_\ell)$ is a permutation of $(x_v, x_w, y^1_{e,j}, \ldots, y^{\ell-3}_{e,j}, y^0_{e,j})$ for some edge $e = (v, w)$ and some $j \in \{1, \ldots, r\}$, except that, in that case that $\mathcal{L}$ is the equation $-x_1 + x_2 + y_1 = y_0$, the sequence (z_1, z_2, z_3, z_4) may also be a permutation of $(y^1_{e,j}, y^0_{e,j}, y^1_{e,j'}, y^0_{e,j'})$ for any edge e and $j \neq j'$;*

5. $\mathrm{maxabs}(A) = \mathcal{O}(|V|^{2\ell+2} r^{\ell+1})$;

6. $A \subseteq \mathbb{N}$ *unless all the coefficients* $c_1, \ldots, c_\ell$ *have the same sign.*

Proof. The proof very largely follows that of Lemma 3.2. As before we have a variable X_v corresponding to each vertex value x_v and a variable $Y^i_{e,j}$ corresponding to each edge value $y^i_{e,j}$; the set of all of these variables is B.

As before, consider any sequence $z = (W_1, W_2, Z_1, \ldots, Z_{\ell-3}, Z_0) \in B^\ell$. We need to show that if $\mathrm{rlc}(z)$ is identically zero, then either z is an edge-sequence or $\mathrm{lc}(z)$ is identically zero. First suppose that $\ell \geq 5$. The only extra case to consider is where there are two terms in the sequence which are dependent edge variables for the same edge e, but from different sets, that is, two variables $Y^0_{e,j}, Y^0_{e,j'}$, with $j \neq j'$, both of which have non-zero total coefficient. But then each requires the other terms $Y^i_{e,j}, Y^i_{e,j'}$ for each $i \neq 0$ to occur as well, giving at least $2(\ell-2)$ terms in all, which is impossible if $\ell \geq 5$ since then $2(\ell-2) > \ell$.

The case $\ell = 4$ is a little more complicated, because in this case there could be further non-trivial solutions in one case only. Consider any sequence $z = (W_1, W_2, Z_1, Z_0) \in B^4$, and suppose that $\mathrm{rlc}(z)$ is identically zero. Then $\mathrm{lc}(z) = a_1 W_1 + a_2 W_2 + b_1 Z_1 - b_0 Z_0$. As above the only way this can happen, apart from an edge solution or a trivial solution, is if two of the variables are $Y^0_{e,j}, Y^0_{e,j'}$, with $j \neq j'$, in which case the other two must be $Y^1_{e,j}, Y^1_{e,j'}$. Thus, for some permutation (p_1, p_2, p_3, p_4) of $(a_1, a_2, b_1, -b_0)$, we have

$$\mathrm{lc}(z) = p_1 Y^0_{e,j} + p_2 Y^1_{e,j} + p_3 Y^0_{e,j'} + p_4 Y^1_{e,j'}. \tag{5}$$

If $e = (v, w)$, where $v < w$, then we know that these variables must satisfy

$$\begin{aligned} a_1 X_v + a_2 X_w + b_1 Y^1_{e,j} &= b_0 Y^0_{e,j} \\ a_1 X_v + a_2 X_w + b_1 Y^1_{e,j'} &= b_0 Y^0_{e,j'}. \end{aligned}$$

Hence substituting for $Y^0_{e,j}$ and $Y^0_{e,j'}$ in (5), we find that

$$\begin{aligned} \mathrm{rlc}(z) = p_1 &\left(\frac{a_1}{b_0} X_v + \frac{a_2}{b_0} X_w + \frac{b_1}{b_0} Y^1_{e,j} \right) + p_2 Y^1_{e,j} \\ &+ p_3 \left(\frac{a_1}{b_0} X_v + \frac{a_2}{b_0} X_w + \frac{b_1}{b_0} Y^1_{e,j'} \right) + p_4 Y^1_{e,j'}. \end{aligned} \tag{6}$$

Since this is identically zero, we must have $p_1 + p_3 = 0$, and so since $p_1 \frac{b_1}{b_0} + p_2 = 0$ and $p_3 \frac{b_1}{b_0} + p_4 = 0$, we also have $p_2 + p_4 = 0$. Hence the coefficients are $\pm p, \pm q$ for some $p \geq q \geq 0$, and so the standard form is

$$-qx_1 + qx_2 + py_1 = py_0.$$

Hence we have $b_1 = b_0 = p$, so that $p_1 \frac{b_1}{b_0} + p_2 = 0$ reduces to $p_1 + p_2 = 0$. Hence all the coefficients are $\pm p_1$, so $p = q$ and so (after dividing by p), the equation in standard form is

$$-x_1 + x_2 + y_1 = y_0.$$

Thus in this case only, we have $\mathrm{lc}(z)$ is not identically zero, and so the values $y^0_{e,j}, y^1_{e,j}, y^0_{e,j'}, y^1_{e,j'}$ for any $j \neq j'$, give further non-trivial solutions to $\mathcal{L}$ in A. □

The problem #Independent Set is defined as follows.

#-Independent Set *Input:* A graph G. *Output:* The number of independent sets of G.

When stated in terms of logical variables and clauses, this problem is better known as #Monotone 2-SAT. The equivalence of the two is demonstrated by replacing logical variables by vertices and clauses by edges. Then satisfying assignments correspond to independent sets via the correspondence mapping the set of variables assigned false in a truth assignment to the corresponding subset of the vertices. It was shown to be #P-complete by Valiant [9].

We shall need the following simple lemma similar to part of Fact 6 from [9].

Lemma 6.3. *Let $z_0, \ldots, z_m$ be non-negative integers and let $p > \sum_{t=0}^{m} z_t$ with $p \in \mathbb{Q}$. Then for each i, z_i is uniquely determined by $\sum_{t=0}^{m} z_t p^t$ and $z_{i+1}, \ldots, z_m$, and may be found in polynomial time.*

Proof. If $\sum_{t=0}^{m} z_t p^t$ and $z_{i+1}, \ldots, z_m$ are known, then one may compute $\sum_{t=0}^{i} z_t p^t$. Moreover $p^i > p^{i-1} \sum_{t=0}^{m} z_t \geq \sum_{t=0}^{i-1} z_t p^t$. Hence $z_i = \lfloor \sum_{t=0}^{i} z_t p^t / p^i \rfloor$. □

Now we can show #$\mathcal{L}$-Free Subset is #P-complete for any homogeneous equation with at least three variables. Again there are two cases, using each of the constructions above.

Theorem 6.4. *Let $\mathcal{L}$ be the linear equation $c_1x_1 + c_2x_2 + c_3x_3 = 0$, where the coefficients c_i are all non-zero. Then #$\mathcal{L}$-Free Subset is #P-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$.*

Proof. We reduce from #Independent Set. Let $G = (V, E)$ be an instance of #Independent Set. Let $m = |E|$, $n = |V|$ and r be the smallest integer such that

$$r > (m + n)\frac{\ln 2}{\ln(4/3)}.$$

We construct, in polynomial time, an instance of #$\mathcal{L}$-Free Subset by applying Lemma 6.1. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the set given by Lemma 6.1 is a subset of $\mathbb{N}$. Write $A = A_E \cup A_V \cup U$, using the same notation as Lemma 6.1. We shall count the number of ways that an $\mathcal{L}$-free subset S of $A_E \cup A_V$ may be extended to an $\mathcal{L}$-free subset of A. Notice that the only solutions to $\mathcal{L}$ in A involving either $u_{v,e,i}$ or $u_{w,e,i}$ include y_e and both $u_{v,e,i}$ and $u_{w,e,i}$. Thus if $y_e \in S$, then one or other but not both of $u_{v,e,i}$ and $u_{w,e,i}$ may be added to S while maintaining $\mathcal{L}$-freeness. If $y_e \notin S$, then any subset of $\{u_{v,e,i}, u_{w,e,i}\}$ may be added to S while maintaining $\mathcal{L}$-freeness. Suppose that $|S \cap A_E| = t$. Then S may be extended to an $\mathcal{L}$-free subset of A in $3^{rt}4^{r(m-t)}$ ways. Let z_t denote the number of $\mathcal{L}$-free subsets of $A_E \cup A_V$ with $|S \cap A_E| = m - t$. Then the number of $\mathcal{L}$-free subsets of A is

$$\sum_{t=0}^{m} 3^{rt}4^{r(m-t)}z_{m-t} = \sum_{t=0}^{m} 3^{rm}(4/3)^{rt}z_t.$$

Now let S be any subset of $A_E \cup A_V$. If $A_E \subseteq S$, then S is $\mathcal{L}$-free if and only if $S \cap A_V$ is an independent set of G. Thus the number of independent sets of G equals z_0. Lemma 6.3

implies that providing $(4/3)^r > \sum_{t=0}^{m} z_t$, z_0 may be found from $\sum_{t=0}^{m}(4/3)^{rt} z_t$ which itself is easily computed from the number of $\mathcal{L}$-free subsets of A. As $\sum_{t=0}^{m} z_t \leq 2^{|A_E|+|A_V|} = 2^{m+n}$, it is sufficient to choose r so that $r > (m+n)\frac{\ln 2}{\ln(4/3)}$. □

Theorem 6.5. *Let $\mathcal{L}$ be the linear equation $c_1 x_1 + \ldots + c_\ell x_\ell = 0$, where $\ell \geq 4$ and the coefficients c_i are all non-zero. Then #$\mathcal{L}$-Free Subset is #P-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$.*

Proof. We reduce from #Independent Set. Let $G = (V, E)$ be an instance of #Independent Set. Let $m = |E|$, $n = |V|$. We construct, in polynomial time, an instance of #$\mathcal{L}$-Free Subset by applying Lemma 6.2. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the set given by Lemma 6.2 is a subset of $\mathbb{N}$. Write $A = A_V \cup A_E$, using the same notation as Lemma 6.2. We shall count the number of ways that an $\mathcal{L}$-free subset S of A_V may be extended to an $\mathcal{L}$-free subset of A. We will say that a subset S of A_V contains an edge $e = (v, w)$ of G if $x_v, x_w \in S$.

We first deal with the case where $\mathcal{L}$ is not equivalent to the equation $x_1 + x_2 = x_3 + x_4$. Consider a subset $S \subseteq A_V$ which contains t edges. Then for each such edge $e \in S$, we can add any set of edge values $y_{e,j}^i$ to S provided that we do not add a full set $\{y_{e,j}^1, \ldots, y_{e,j}^{\ell-3}, y_{e,j}^0\}$ for any j, hence there is a choice of $(2^{\ell-2}-1)^r$ subsets for each of these t edges. On the other hand, for any edge e not in S, we can add any subset of the edge values for e, hence there is a choice of $(2^{\ell-2})^r$ subsets. Thus if z_t is the number of subsets of A_V containing $m - t$ edges, then the number of $\mathcal{L}$-free subsets of A is

$$\sum_{t=0}^{m}(2^{\ell-2}-1)^{r(m-t)}(2^{\ell-2})^{rt} z_t = (2^{\ell-2}-1)^{rm} \sum_{t=0}^{m} z_t \left(\frac{2^{\ell-2}}{2^{\ell-2}-1}\right)^{rt}.$$

Hence if r is the smallest integer such that

$$r > \frac{n \ln 2}{\ln(2^{\ell-2}) - \ln(2^{\ell-2}-1)},$$

then by Lemma 6.3, we can determine z_m from this. But z_m is the number of subsets of A_V containing no edges, that is, the number of independent sets of G, as required.

Now consider the special case of the equation $-x_1 + x_2 + y_1 = y_0$. Again consider a subset $S \subseteq A_V$ which contains t edges. As before, for each of these edges, we can add any set of edge values $y_{e,j}^i$ to S provided that we do not add a full set $\{y_{e,j}^1, y_{e,j}^0\}$ for any j, hence there is a choice of 3^r subsets for each of these t edges.

For any edge e not in S, we can add all of these sets, but we could also add any set which includes $\{y_{e,j}^1, y_{e,j}^0\}$ for exactly one value of j. Hence in total we can add $3^r + r.3^{r-1}$ subsets. No other sets are possible because these would contain the special extra non-trivial solutions to $\mathcal{L}$. Thus as before, if z_t is the number of subsets of A_V containing $m - t$ edges, then the number of $\mathcal{L}$-free subsets of A is

$$\sum_{t=0}^{m}(3^r)^{m-t}((r+3)3^{r-1})^t z_t = 3^{mr} \sum_{t=0}^{m} \left(\frac{r+3}{3}\right)^t z_t.$$

Hence if we can determine this number for $m+1$ different values of r, we have a system of $m+1$ equations in the quantities $z_0, \ldots, z_m$. Since the coefficients have a non-zero Vandermonde determinant, the solution is unique and can be determined in polynomial time in the total size of the coefficients [3]. Hence we can determine z_m as before. □

We now consider the counting version of $\mathcal{L}$-Free Subset when $\mathcal{L}$ is an inhomogeneous equation. We shall need a final variant of Lemma 3.2.

Lemma 6.6. *Consider a linear equation $c_1x_1 + \cdots + c_\ell x_\ell = K$, where $\ell \geq 3$ and the coefficients c_i and constant K are all non-zero, with $\gcd(c_1, \ldots, c_\ell)$ a divisor of K. Let $G = (V, E)$ be a bipartite graph and let $r \in \mathbb{N}$. Then we can construct in polynomial time, a set $A \subseteq \mathbb{Z}$ with the following properties.*

1. *A is the union of ℓ sets A_V, A_E^0, $A_E^1, \ldots, A_E^{\ell-3}$ and U, where $A_V = \{x_v : v \in V\}$, $A_E^i = \{y_e^i : e \in E\}$ for each $i = 0, 1, \ldots, \ell-3$ and $U = \{u_{v,e,i} : v \in V, e = vw \in E, 1 \leq i \leq r\}$;*
2. *$|A| = |V| + (\ell - 2 + 2r)|E|$;*
3. *for every edge $e = vw$, some permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ is a solution to $\mathcal{L}$;*
4. *for every edge $e = vw$ and every integer i with $1 \leq i \leq r$, some permutation of $(u_{v,e,i}, u_{w,e,i}, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ is a solution to $\mathcal{L}$;*
5. *if $(z_1, \ldots, z_\ell)$ is a solution to $\mathcal{L}$, then $(z_1, \ldots, z_\ell)$ is a permutation of $(x_v, x_w, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ for some edge $e = vw$ or of $(u_{v,e,i}, u_{w,e,i}, y_e^1, \ldots, y_e^{\ell-3}, y_e^0)$ for some edge $e = vw$ and integer i with $1 \leq i \leq r$.*
6. *$\operatorname{maxabs}(A) = \mathcal{O}(|V|^{2\ell+2}r)$;*
7. *$A \subseteq \mathbb{N}$ unless all the coefficients $c_1, \ldots, c_\ell$ have the same sign.*

Proof. Apply Lemma 3.8, by regarding G as a tripartite graph with $V_3 = \emptyset$, giving sets A_V and A_E as described in Lemma 3.8. Let $c = \max\{|c_1|, \ldots, |c_\ell|\}$ and $N = 2\ell c \operatorname{maxabs}(A_E \cup A_V) + 1$.

Next for each e in E and integer i with $1 \leq i \leq r$ choose a strictly positive integer $N_{e,i}$ so that if $(e_1, i_1) \neq (e_2, i_2)$ and $j_1 \neq j_2$, then the following holds

$$c_{j_1}c_1N_{e_1,i_1} - c_{j_2}c_2N_{e_2,i_2} \neq 0. \tag{7}$$

In particular, the integers $N_{e,i}$ are pairwise distinct. We may choose the integers $N_{e,i}$ greedily in any order. When one of them is chosen, there are strictly fewer than $2|E|r\ell^2$ linear equations that it must not satisfy. Hence one may choose these integers from $\{1, \ldots, 2|E|r\ell^2\}$.

We now construct the set U. Start with U being empty, and for each edge of G, add $2r$ elements to it. Suppose that the edge e of G joins v and w, where $v \in V_1$ and $w \in V_2$ in the notation of Lemma 3.8. For each $i = 1, \ldots, r$, add the elements $u_{v,e,i} = x_v - NN_{e,i}c_2$ and $u_{w,e,i} = x_w + NN_{e,i}c_1$ to U.

Notice that all the conditions on A in the statement of the lemma are satisfied except possibly Conditions 5 and 7. If the coefficients of $\mathcal{L}$ do not all have the same sign, then we may assume that c_1 is positive and c_2 is negative. This ensures that Condition 7 is satisfied. We now show that Condition 5 is also satisfied.

Given an ℓ-tuple $(z_1, \ldots, z_\ell)$, let $\mathcal{L}(z_1, \ldots, z_\ell) = \sum_{i=1}^{\ell} c_i z_i$. If $z \in A$, then define $\bar{z} = z$ to be the unique element of $A \backslash U$ to which it is congruent modulo N. Suppose that $(z_1, \ldots, z_\ell)$ is an ℓ-tuple of elements of A satisfying $\mathcal{L}(z_1, \ldots, z_\ell) = K$. Then $\mathcal{L}(\bar{z}_1, \ldots, \bar{z}_\ell) \equiv K$

(mod N). But $|\mathcal{L}(\bar{z}_1, \ldots, \bar{z}_\ell)| < N/2$, so $\mathcal{L}(\bar{z}_1, \ldots, \bar{z}_\ell) = K$ and $(\bar{z}_1, \ldots, \bar{z}_\ell)$ is a permutation of an edge solution to $\mathcal{L}$. Exactly two elements of $\{\bar{z}_1, \ldots, \bar{z}_\ell\}$ belong to A_V. Suppose without loss of generality that $\bar{z}_r = x_v$ and $\bar{z}_s = x_w$, with $v \in V_1$ and $w \in V_2$ in the notation of Lemma 3.8. Then $z_r = x_v$ and $z_s = x_w$, or $z_r = x_v - NN_{e_1,i_1}c_2$ and $z_s = x_w + NN_{e_2,i_2}c_1$ for some e_1, e_2, i_1, i_2. In the latter case, the choice of the integers $N_{e,i}$ ensures that $e_1 = e_2 = vw$ and $i_1 = i_2$. □

Theorem 6.7. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \ldots + c_\ell x_\ell = K$, where $\ell \geq 3$ and K and the coefficients c_i are all non-zero. Then #$\mathcal{L}$-Free Subset is #P-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$.*

Proof. The proof is very similar to that of Theorem 6.4, except that this time we apply Lemma 6.6 and use the fact that #Independent Set remains #P-hard when the input is restricted to being a bipartite graph [7]. □

Combining Theorems 6.4, 6.5 and 6.7 gives the following:

Theorem 6.8. *Let $\mathcal{L}$ be a linear equation $c_1x_1 + \ldots + c_\ell x_\ell = K$, where $\ell \geq 3$ and the coefficients c_i are all non-zero. Then #$\mathcal{L}$-Free Subset is #P-complete. If the coefficients of $\mathcal{L}$ are not all of the same sign, then the input set A can be restricted to be a subset of $\mathbb{N}$.*

References

[1] P. Alimonti and V. Kann, Some APX-completeness results for cubic graphs, Theoret. Comput. Sci. **237** (2000), no. 1-2, 123–134.

[2] S. Eberhard, B. Green and F. Manners, Sets of integers with no large sum-free subset, Ann. of Math. (2) **180** (2014), no. 2, 621–652.

[3] J. Edmonds, Systems of distinct representatives and linear algebra, J. Res. Nat. Bur. Standards Sect. B **71B** (1967), 241–245.

[4] P. Erdős, Extremal problems in number theory, Proc. Sympos. Pure Math., Vol. VIII, Amer. Math. Soc., Providence, R.I. (1965), 181–189.

[5] M. R. Garey, D. S. Johnson and L. Stockmeyer, Some simplified NP-complete graph problems, Theoret. Comput. Sci. **1** (1976), no. 3, 237–267.

[6] K. Meeks and A. Treglown, On the complexity of finding and counting solution-free sets of integers, Discrete Appl. Math. **243** (2018) 219–238.

[7] J. S. Provan and M. O. Ball. The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM Journal on Computing*, **12** (1983) 777–788.

[8] K.F. Roth, On certain sets of integers, *J. London Math. Soc.*, **28**, (1953), 104–109.

[9] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, **8** (1979) 410–421.