BOUNDS FOR THE LOCAL PROPERTIES PROBLEM FOR DIFFERENCE SETS

SANJANA DAS

ABSTRACT. We consider the local properties problem for difference sets: we define $g(n, k, \ell)$ to be the minimum value of $|A - A|$ over all n-element sets $A \subseteq \mathbb{R}$ with the 'local property' that $|A' - A'| \geq \ell$ for all k-element subsets $A' \subseteq A$. We view k and ℓ as fixed, and study the asymptotic behavior of $g(n, k, \ell)$ as $n \to \infty$. One of our main results concerns the quadratic threshold, i.e., the minimum value of ℓ such that $g(n, k, \ell) = \Omega(n^2)$; we determine this value exactly for even k, and we determine it up to an additive constant for odd k. We also show that for all $1 < c \leq 2$, the 'threshold' for $g(n, k, \ell) = \Omega(n^c)$ is quadratic in k; conversely, for ℓ quadratic in k, we obtain upper and lower bounds of the form n^c for (not necessarily equal) constants $c > 1$. In particular, this provides the first nontrivial upper bounds in the regime where ℓ is quadratic in k.

1. INTRODUCTION

1.1. **History of local properties problems.** There is a long history of studying objects that have local properties, and trying to understand the extent to which these local properties determine global properties. In particular, many questions of this form concern going from local to global structure: if we know every small piece of an object is 'unstructured' in some sense, how unstructured does the *entire* object have to be?

For example, Erdős and Shelah [3, Section V] initiated the study of a local properties problem for graphs, where we consider edge-colorings of a complete graph and think of an edge-coloring as 'unstructured' if it contains many distinct colors. For positive integers n, k, and ℓ, we define $f(n, k, \ell)$ to be the minimum number of colors in an edge-coloring of K_n with the property that every induced subgraph K_k contains at least ℓ colors. We think of k and ℓ as fixed and n as large, and we wish to understand the asymptotic behavior of $f(n, k, \ell)$ as n tends to infinity. There are two natural ways to approach this problem. One is to fix k and ℓ and attempt to bound $f(n, k, \ell)$. The other is to attempt to understand various thresholds for $f(n, k, \ell)$ — given k, we search for the smallest value of ℓ (as a function of k) for which we have a certain lower bound on $f(n, k, \ell)$.

This local properties problem for graphs has been studied by many authors: Erdős and Gyárfás [4] established the linear and quadratic thresholds (the smallest values of ℓ as a function of k for which $f(n, k, \ell) = \Omega(n)$ and $f(n, k, \ell) = \Omega(n^2)$, respectively) and an upper bound on the polynomial threshold (the smallest value of ℓ for which $f(n, k, \ell) = \Omega(n^\varepsilon)$ for some $\varepsilon > 0$); Conlon, Fox, Lee, and Sudakov [2] later proved a matching lower bound, thus establishing the polynomial threshold. Erdős and Gyárfás [4] also proved a general upper bound for $f(n, k, \ell)$ using a probabilistic construction. Pohoata and Sheffer [8] and Fish, Pohoata, and Sheffer [6] later proved lower bounds for $f(n, k, \ell)$ in several regimes, which grow arbitrarily close to the upper bound of Erdős and Gyárfás as k grows large. A more detailed discussion of this problem can be found in [6] and [8].

Erdős [3] also posed a similar local properties problem for distinct distances, to determine the minimum number of distinct distances spanned by a set of n points in $\mathbb{R}^2$ such that every k points span at least ℓ distinct distances. This problem has also been studied by various authors; more discussion of this problem can be found in [5].

1.2. **A local properties problem for difference sets.** We study an arithmetic version of such local properties problems, suggested by Zeev Dvir (see [8]). We consider sets of real numbers for which every small piece is 'arithmetically unstructured,' and we wish to understand how arithmetically unstructured this local property requires the entire set to be. One measure of the extent to which a set is arithmetically structured is the size of its difference set — roughly, sets with smaller difference set relative to their size

possess more arithmetic structure. So we consider sets for which every small subset has a reasonably large difference set, and attempt to understand how large the difference set of the entire set must be.

We now make this precise. For a set $A \subseteq \mathbb{R}$, we define the *difference set* of A to be

$$A - A = \{|a - b| \mid a, b \in A, \ a \neq b\}.$$

Note that we include only positive differences in $A - A$; this definition is somewhat nonstandard but more natural for the problem we study. (When discussing the differences present in a set, we refer only to positive differences, unless otherwise stated.)

For integers $n \geq k \geq 1$ and $\ell \geq 0$, we define $g(n, k, \ell)$ to be the minimum value of $|A - A|$ over all n-element sets $A \subseteq \mathbb{R}$ with the property that every k-element subset $A' \subseteq A$ satisfies $|A' - A'| \geq \ell$; we refer to this property as the (k, ℓ)-*local property*. (Note that the same definition can be used to define $g(n, k, \ell)$ even when ℓ is not a nonnegative integer; then $g(n, k, \ell) = g(n, k, \lceil \ell \rceil)$. This makes certain results slightly more convenient to state.) We think of k and ℓ as fixed and n as large, and we study the asymptotic behavior of $g(n, k, \ell)$ as n tends to infinity. (In particular, in all asymptotic notation in this paper, n tends to infinity while all other quantities are fixed; the implied constants may depend on k, ℓ, and any other stated parameters.)

As with the local properties problem for graph colorings, we can attempt to understand $g(n, k, \ell)$ either by fixing ℓ, or by considering various thresholds. Note that for all n-element sets $A \subseteq \mathbb{R}$ we have $n - 1 \leq |A - A| \leq \binom{n}{2}$ (and $|A - A| = n - 1$ if and only if A is an arithmetic progression), so for all $\ell \leq \binom{k}{2}$ we have

$$n - 1 \leq g(n, k, \ell) \leq \binom{n}{2}.$$

(If $\ell > \binom{k}{2}$, then no set A can satisfy the (k, ℓ)-local property, so we say that $g(n, k, \ell) = +\infty$.) Also note that $g(n, k, \ell)$ is weakly increasing in ℓ, as increasing ℓ only makes the (k, ℓ)-local property stronger. So it is natural to consider, for fixed k, how long $g(n, k, \ell)$ stays linear and when it becomes quadratic as we increase ℓ. To formalize this, as defined in [7], the *superlinear threshold* is defined to be the largest integer ℓ (as a function of k) such that $g(n, k, \ell) = O(n)$, and the *quadratic threshold* is defined to be the smallest integer ℓ (as a function of k) such that $g(n, k, \ell) = \Omega(n^2)$. (Note that $g(n, k, k - 1) = n - 1$, so at the superlinear threshold we actually have $g(n, k, \ell) = \Theta(n)$; similarly, if $k \geq 4$ then $g(n, k, \binom{k}{2}) = \binom{n}{2}$, so at the quadratic threshold we actually have $g(n, k, \ell) = \Theta(n^2)$.)

1.3. **Previous bounds.** This problem was first studied by Fish, Pohoata, and Sheffer [6], who proved the following family of lower bounds: for all $r \geq 2$ and k divisible by $2r$, we have

$$(1) \qquad g\left(n, k, \binom{k}{2} - \frac{(r-1)(r+2)}{2}\binom{k/r}{2} + 1\right) = \Omega(n^{\frac{r}{r-1} \cdot \frac{k-2r}{k}}).$$

For example, taking $r = 2$, this gives that for $4 \mid k$ we have $g(n, k, \frac{k^2}{4} + 1) = \Omega(n^{2 - \frac{8}{k}})$; more generally, if we think of r as fixed and k as large, this provides a family of lower bounds for $g(n, k, \ell)$ for values of ℓ that are quadratic in k, with leading coefficient $\frac{1}{2} - \frac{(r-1)(r+2)}{4r^2}$ (which ranges from $\frac{7}{32}$ to $\frac{1}{4}$). Fish, Pohoata, and Sheffer obtained these bounds using the technique of color energy, which they had developed to study the local properties problem for graph colorings.

Fish, Pohoata, and Sheffer also proved the following upper bound for $g(n, k, \ell)$ (for certain values of ℓ): for any $\varepsilon > 0$, there exists $c > 0$ such that for all sufficiently large k we have

$$g(n, k, ck(\log k)^{\frac{1}{4} - \varepsilon}) = n \cdot 2^{O(\sqrt{n \log n})}.$$

The idea of their proof is to consider 3-term arithmetic progressions — results from additive combinatorics give that for such values of k and ℓ, any k-element set with fewer than ℓ differences must contain a 3-term arithmetic progression. Then any set A with no 3-term arithmetic progressions satisfies the (k, ℓ)-local property, and Behrend's construction in [1] provides such sets A with reasonably small $|A - A|$.

Fish, Lund, and Sheffer [5] then proved the upper bound

$$g\left(n, k, \frac{k^{\log_2 3} - 1}{2}\right) = O(n^{\log_2 3}).$$

Their construction for the set A was roughly an affine t-cube (i.e., a set of the form $\{a + \varepsilon_1 d_1 + \cdots + \varepsilon_t d_t \mid \varepsilon_1, \ldots, \varepsilon_t \in \{0, 1\}\}$ for fixed $a, d_1, \ldots, d_t$) with $t \approx \log_2 n$.

This problem was then studied by Li [7], who proved that the superlinear threshold is $k-1$ by showing that $g(n, k, k) = \omega(n)$, i.e., if A is an n-element set with no k-term arithmetic progression, then we must have $|A - A| = \omega(n)$.

Li also proved several lower bounds for $g(n, k, \ell)$: if k is even, then

$$(2) \qquad g\left(n, k, \frac{3k^2}{8} - \frac{3k}{4} + 2\right) = \Omega(n^2);$$

if k is divisible by 8, then

$$(3) \qquad g\left(n, k, \frac{9k^2}{32} - \frac{9k}{16} + 5\right) = \Omega(n^{4/3});$$

and if $k \geq 8$ is a power of 2, then

$$(4) \qquad g\left(n, k, \frac{k^{\log_2 3} + 1}{2}\right) = \Omega(n^{1 + \frac{2}{k-2}}).$$

In particular, (2) implies that the quadratic threshold is at most $\frac{3k^2}{8} - \frac{3k}{4} + 2$. The proofs of all three of these bounds can be interpreted as showing that a set with sufficiently small difference set must contain $\frac{k}{2^t}$ disjoint congruent affine t-cubes (where $t = 1$, 2, and $\log_2 k - 1$, respectively), whose elements form a k-element subset contradicting the (k, ℓ)-local property for the stated values of ℓ.

Li also proved, using a random construction, that for all $c \geq 2$ and k large with respect to c, we have

$$g(n, k, ck + 1) = O(n^{1 + \frac{c^2 + 1}{k}}).$$

1.4. **Our results.** In this paper, we prove a family of upper and lower bounds for $g(n, k, \ell)$. These bounds have several consequences that allow us to better understand the behavior of $g(n, k, \ell)$; we state these consequences first, as they are more intuitive.

First, we establish the quadratic threshold for all even k.

Corollary 1.1. *For all even k, we have*

$$g\left(n, k, \frac{k^2}{4}\right) = o(n^2) \ \text{and} \ g\left(n, k, \frac{k^2}{4} + 1\right) = \Omega(n^2).$$

In particular, the quadratic threshold is $\frac{k^2}{4} + 1$.

For odd k, we are not able to determine the exact value of the quadratic threshold, but we obtain upper and lower bounds for the quadratic threshold that differ by 3.

Corollary 1.2. *For all odd k, we have*

$$g\left(n, k, \frac{(k+1)^2}{4} - 4\right) = o(n^2) \ \text{and} \ g\left(n, k, \frac{(k+1)^2}{4}\right) = \Omega(n^2).$$

In particular, the quadratic threshold is between $\frac{(k+1)^2}{4} - 3$ and $\frac{(k+1)^2}{4}$, inclusive.

We also obtain bounds on the thresholds for other exponents, although these bounds are less tight — for each $1 < c \leq 2$, we define the *threshold for $\Omega(n^c)$* to be the smallest integer ℓ such that $g(n, k, \ell) = \Omega(n^c)$. (The quadratic threshold is the special case of this definition where $c = 2$.) We obtain that for each fixed c, the threshold for $\Omega(n^c)$ is quadratic in k.

Corollary 1.3. *For each $1 < c \leq 2$, there exist constants $0 < a_1 < a_2 < \frac{1}{2}$ — which can be defined as explicit functions of c such that $a_1, a_2 \to 0$ as $c \to 1$ — such that the following holds for all sufficiently large k: we have $g(n, k, a_1 k^2) = o(n^c)$ and $g(n, k, a_2 k^2) = \Omega(n^c)$. In particular, the threshold for $\Omega(n^c)$ is between $a_1 k^2$ and $a_2 k^2$.*

We can also view our results in terms of bounds on $g(n, k, \ell)$ for given ℓ — when ℓ is quadratic in k with leading coefficient less than $\frac{1}{4}$, we obtain both upper and lower bounds on $g(n, k, \ell)$ of the form n^c for $1 < c < 2$ (though not with the same exponent). In particular, these bounds provide the first nontrivial upper bounds on $g(n, k, \ell)$ for values of ℓ that are quadratic in k (the previously known upper bounds work with much smaller values of ℓ).

Corollary 1.4. *For each $0 < a < \frac{1}{4}$, there exist $1 < c_1 < c_2 < 2$ — which can be defined as explicit functions of a such that $c_1, c_2 \to 1$ as $a \to 0$ — such that for all sufficiently large k, we have $g(n, k, ak^2) = \Omega(n^{c_1})$ and $g(n, k, ak^2) = o(n^{c_2})$.*

Finally, for each k, one can consider the set of 'possible exponents' that occur in the asymptotics for $g(n, k, \ell)$ over all ℓ — to make this precise, we define

$$S_k = \left\{ \liminf_{n \to \infty} \frac{\log g(n, k, \ell)}{\log n} \; \middle| \; 0 \le \ell \le \binom{k}{2} \right\}.$$

As of now, very little is known about this set, beyond the fact that it contains 1 for all k and 2 for all $k \ge 4$. However, we are able to show that $|S_k|$ grows arbitrarily large as k does.

Corollary 1.5. *There exists an absolute constant $a > 0$ such that for all $k \ge 2$, we have $|S_k| \ge a \log \log k$.*

We now state our more general theorems from which we deduce these statements. First, we prove the following family of upper bounds for $g(n, k, \ell)$.

Theorem 1.6. *For all positive integers k and all $1 < c \le 2$, we have*

$$g\left(n, k, \left\lceil \frac{(c-1)(k-1)}{c} \right\rceil^2\right) = o(n^c).$$

In order to obtain our lower bound on the quadratic threshold for odd k (as stated in Corollary 1.2), we also prove the following improvement on Theorem 1.6 in the case where $c = 2$ and k is odd.

Proposition 1.7. *For all odd positive integers k, we have*

$$g\left(n, k, \frac{(k+1)^2}{4} - 4\right) = o(n^2).$$

We obtain Theorem 1.6 and Proposition 1.7 by using a random construction. To ensure that the constructed set satisfies the (k, ℓ)-local property for the stated values of ℓ, we consider all possible 'configurations' of equal differences that can be formed by a k-element subset. We first show, very roughly, that certain 'bad' configurations are expected to occur very few times in the random construction, so using the alteration method, we can ensure that our set only contains 'good' configurations. We then show that any k-element subset forming a good configuration must contain at least ℓ distinct differences. The technical steps of our proof rely on considering the systems of equations associated with such configurations and analyzing their linear algebraic properties.

We also prove the following family of lower bounds for $g(n, k, \ell)$.

Theorem 1.8. *For all positive integers t and k with $2^t \mid k$, we have*

$$g\left(n, k, \frac{3^{t-1}}{4^t} k^2 + \frac{3^{t-1}+1}{2}\right) = \Omega(n^{1 + \frac{1}{2^t - 1}}).$$

Using certain straightforward monotonicity properties of $g(n, k, \ell)$, we can extend Theorem 1.8 to all values of k (not just multiples of 2^t), with an adjustment in the value of ℓ that is linear in k; we state this extension in Corollary 5.5.

We obtain Theorem 1.8 by showing that a set with small difference set must contain a *specific* k-element configuration that has few distinct differences. This idea is similar to Li's proofs of (2), (3), and (4); instead of using a configuration consisting of $\frac{k}{2^t}$ congruent affine t-cubes, we use one consisting of $\frac{k}{2^{t-1}}$ congruent affine $(t-1)$-cubes whose centers form $\frac{k}{2^t}$ pairs with equal sums. (In particular, to obtain the lower bound of $\Omega(n^2)$, we use a configuration consisting of $\frac{k}{2}$ pairs with equal sums instead of $\frac{k}{2}$ pairs with equal differences.) If we think of t as fixed and k as large, this improves the value of ℓ in the bounds by a factor of around $\frac{2}{3}$.

If we consider c (in Theorem 1.6) and t (in Theorem 1.8) to be fixed and k to be large, then Theorems 1.6 and 1.8 provide upper and lower bounds on $g(n, k, \ell)$ with constant exponents of n (not depending on k), for values of ℓ that are quadratic in k (with coefficients depending on the exponent in the bound). The relationships between the leading coefficient in ℓ and the exponent of n in our two bounds are illustrated in Figure 1, with the blue line representing the upper bounds and the purple line representing the lower bounds. Although there is a gap between the two lines below the point $(\frac{1}{4}, 2)$, both approach the point $(0, 1)$

— for the blue line, this is because we are able to obtain an upper bound of n^c for every $1 < c \leq 2$, and the corresponding leading coefficient $(\frac{c-1}{c})^2$ tends to 0 as $c \to 1$; for the purple line, this is because in our lower bounds, the leading coefficient $\frac{3^{t-1}}{4^t}$ tends to 0 as $t \to \infty$, and the corresponding exponent $1 + \frac{1}{2^t-1}$ tends to 1. Very roughly, this feature of the graph, along with the fact that such a graph can be made in the first place, is what allows us to deduce Corollaries 1.3 and 1.4.
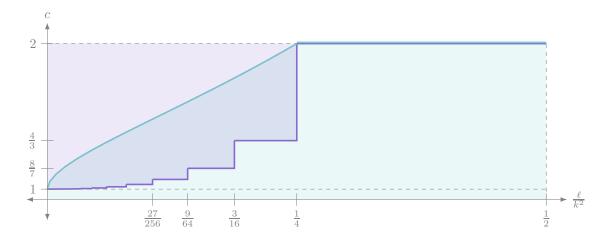

FIGURE 1. This graph depicts the relationship between the coefficients of k^2 in ℓ (on the x-axis) and the exponents c (on the y-axis) for which we can bound $g(n, k, \ell)$ above or below by n^c. The blue line represents the upper bounds in Theorem 1.6, and the purple line represents the lower bounds in Theorem 1.8.

In Section 2, we prove Theorem 1.6. In Section 3, we prove Proposition 1.7, by strengthening the previous proof in the case where $c = 2$ and k is odd. In Section 4, we prove Theorem 1.8. Finally, in Section 5, we deduce Corollaries 1.1, 1.2, 1.3, 1.4, and 1.5.

2. PROOF OF THEOREM 1.6 (FAMILY OF UPPER BOUNDS ON $g(n, k, \ell)$)

In this section, we prove Theorem 1.6 (our family of upper bounds). The structure of the proof is as follows: let ℓ be as in Theorem 1.6. In order to show that $g(n, k, \ell) = o(n^c)$, we need to show that for all $a > 0$, for all sufficiently large n we have $g(n, k, \ell) \leq an^c$ — i.e., there exists an n-element set $A \subseteq \mathbb{R}$ with $|A - A| \leq an^c$ satisfying the (k, ℓ)-local property. We will obtain such a set A using a random construction. In order to ensure that our set A satisfies the (k, ℓ)-local property, for every k-element subset $A' \subseteq A$, we consider the 'configuration' of equal differences formed by its elements. Our argument will then consist of two steps: first, we show that using a random construction, we can obtain an n-element set A with $|A - A| \leq an^c$ that avoids all 'bad' configurations (where we classify configurations as good or bad based on certain properties). We then show that any k-element subset forming a 'good' configuration necessarily contains at least ℓ distinct differences. It then follows that the set A we constructed must satisfy the (k, ℓ)-local property, providing the desired upper bound.

In Subsection 2.1, we set up a few useful definitions (in particular, we formalize the notion of a configuration of equal differences) and state these two steps more precisely (with definitions of which configurations we consider good and bad). In Subsection 2.2, we describe our random construction and prove that it avoids all bad configurations. In Subsection 2.3, we prove that any subset forming a good configuration contains at least ℓ distinct differences, modulo a certain technical lemma; in Subsection 2.4, we prove this lemma.

2.1. **Setup.** We formalize the notion of a configuration of equal differences in the following way: we define a *k-configuration* to be a collection of linear equations over $\mathbb{R}$ in the k variables $x_1, \ldots, x_k$, in which each equation is of the form

$$x_{i_1} - x_{i_2} - x_{i_3} + x_{i_4} = 0 \tag{5}$$

for some (not necessarily distinct) indices $i_1, i_2, i_3, i_4 \in \{1, \ldots, k\}$ such that $\{i_1, i_4\} \neq \{i_2, i_3\}$. We refer to such an equation as a *difference equality*. (The name, as well as the reason we care about such equations, comes from the fact that (5) is a rearrangement of $x_{i_1} - x_{i_2} = x_{i_3} - x_{i_4}$. The condition $\{i_1, i_4\} \neq \{i_2, i_3\}$ means that we only consider equations that are not identically satisfied.)

In the remainder of this subsection, we establish some useful definitions regarding k-configurations; using these definitions, we then state our two main steps more precisely, as Lemmas 2.4 and 2.5.

2.1.1. *Conventions.* We first fix a few conventions regarding linear equations (working over the field $\mathbb{R}$).

We consider equations to be unique only up to rearrangement; for example, we consider (5) to be the same equation as $-x_{i_1} + x_{i_2} + x_{i_3} - x_{i_4} = 0$, $x_{i_1} - x_{i_2} = x_{i_3} - x_{i_4}$, and $x_{i_1} + x_{i_4} = x_{i_2} + x_{i_3}$.

Given a linear equation $(*)$ in $x_1, \ldots, x_k$, we define its *content* to be an expression $*$ (which is a linear combination of $x_1, \ldots, x_k$) such that $(*)$ is the equation $* = 0$. (For any nontrivial equation $(*)$, there are two ways to define the content of $(*)$, where each is the negative of the other; we choose one arbitrarily.) For example, the content of the difference equality (5) is either $x_{i_1} - x_{i_2} - x_{i_3} + x_{i_4}$ or its negative.

We say that a variable x_i *appears* in a linear equation $(*)$, or that $(*)$ *involves* x_i, if the coefficient of x_i in $(*)$ is nonzero. (For example, the equation $x_1 - x_2 = 0$ involves x_1 and x_2, but does not involve x_3; this remains the case even if we write it as $x_1 - x_2 - x_3 + x_3 = 0$.) We also use the same terminology for linear expressions $*$ in place of linear equations $(*)$. Given a set $I \subseteq \{1, \ldots, k\}$, we say $(*)$ is an *equation on I* if all variables x_i appearing in $(*)$ satisfy $i \in I$.

We say linear equations $(*_1), \ldots, (*_t)$ are *linearly independent* if their contents $*_1, \ldots, *_t$ are linearly independent (when viewed as elements of the k-dimensional $\mathbb{R}$-vector space with basis vectors $x_1, \ldots, x_k$).

Given linear equations $(*_1), \ldots, (*_t), (*)$ with contents $*_1, \ldots, *_t, *$, we say that the collection $\mathcal{T} = \{(*_1), \ldots, (*_t)\}$ *implies* $(*)$ if $*$ can be written as a linear combination of $*_1, \ldots, *_t$, i.e.,

$$* = \varepsilon_1 *_1 + \cdots + \varepsilon_t *_t$$

for some $\varepsilon_1, \ldots, \varepsilon_t \in \mathbb{R}$. (The reason for this terminology is that this is equivalent to the condition that every solution to $\mathcal{T}$ (viewed as a system of equations) is also a solution to $(*)$; however, we will not need this fact.) We say $\mathcal{T}$ *minimally implies* $(*)$ if $\mathcal{T}$ implies $(*)$, but no proper subset $\mathcal{T}' \subset \mathcal{T}$ does. (Equivalently, $\mathcal{T}$ minimally implies $(*)$ if and only if $(*_1), \ldots, (*_t)$ are linearly independent and the coefficients $\varepsilon_1, \ldots, \varepsilon_t$ above are all nonzero.)

We say two collections of linear equations $\mathcal{C}_1$ and $\mathcal{C}_2$ are *equivalent* if $\mathcal{C}_1$ implies every equation in $\mathcal{C}_2$, and $\mathcal{C}_2$ implies every equation in $\mathcal{C}_1$.

2.1.2. *Definitions for k-configurations.* We now flesh out the concept of k-configurations by defining a few attributes of a k-configuration that will be useful for our analysis.

For a k-configuration $\mathcal{C}$, we let $\mathrm{Sol}(\mathcal{C}) \subseteq \mathbb{R}^k$ denote the set of solutions to $\mathcal{C}$ (where we view $\mathcal{C}$ as a system of linear equations over $\mathbb{R}$). For each $I \subseteq \{1, \ldots, k\}$, we let $\mathrm{Sol}_I(\mathcal{C})$ denote the projection of $\mathrm{Sol}(\mathcal{C})$ onto the coordinates in I — in other words, we define

$$\mathrm{Sol}_I(\mathcal{C}) = \{(x_i)_{i \in I} \mid (x_1, \ldots, x_k) \in \mathrm{Sol}(\mathcal{C})\}.$$

We refer to the elements of $\mathrm{Sol}_I(\mathcal{C})$ as *I-solutions to $\mathcal{C}$*.

We define the *dimension* of $\mathcal{C}$, denoted $\dim(\mathcal{C})$, to be the dimension of $\mathrm{Sol}(\mathcal{C})$ as a $\mathbb{R}$-vector space; similarly, for each $I \subseteq \{1, \ldots, k\}$, we define $\dim_I(\mathcal{C})$ to be the dimension of $\mathrm{Sol}_I(\mathcal{C})$.

We say a solution to $\mathcal{C}$ is *generic* if it does not satisfy any difference equalities other than the ones implied by $\mathcal{C}$. Of course, any solution to $\mathcal{C}$ must satisfy all equations implied by $\mathcal{C}$, so a generic solution to $\mathcal{C}$ satisfies all the difference equalities implied by $\mathcal{C}$ and no others. (This definition will be used to encapsulate what we mean by a k-element set 'forming' a configuration of equal differences, as stated in the outline in the beginning of Section 2.)

We define an *occurrence* of a k-configuration $\mathcal{C}$ in a set $A \subseteq \mathbb{R}$ to be a solution to $\mathcal{C}$ consisting of *distinct* elements of A; if there exists an occurrence of $\mathcal{C}$ in A, we say that $\mathcal{C}$ *occurs* in A. For each $I \subseteq \{1, \ldots, k\}$, we define an *$I$-occurrence* of $\mathcal{C}$ to be an I-solution to $\mathcal{C}$ consisting of distinct elements of A.

Example 2.1. We illustrate these definitions for the 6-configuration

$$\mathcal{C} = \{x_1 - x_2 - x_3 + x_4 = 0,\ x_1 - x_2 - x_5 + x_6 = 0\}.$$

- We have $\dim(\mathcal{C}) = 4$ — to obtain a solution to $\mathcal{C}$, we can choose x_1, x_2, x_3, and x_5 arbitrarily, and then x_4 and x_6 are uniquely determined. For the same reason, we have $\dim_{\{1,2,3,5\}}(\mathcal{C}) = 4$ and $\dim_{\{1,2,3,4\}}(\mathcal{C}) = 3$ (as x_4 is uniquely determined by x_1, x_2, and x_3). Similarly, we have $\dim_{\{3,4,5,6\}}(\mathcal{C}) = 3$ — this is because $\mathcal{C}$ implies $x_3 - x_4 - x_5 + x_6 = 0$, so we can choose x_3, x_4, and x_5 arbitrarily, and then x_6 is uniquely determined.
- The 6-tuple $(1, 2, 4, 5, 9, 10)$ is a solution to $\mathcal{C}$, but it is not generic, as it satisfies the additional difference equality $x_1 - x_4 = x_4 - x_5$. By contrast, $(1, 2, 4, 5, 10, 11)$ is a generic solution to $\mathcal{C}$.
- Consider the set $A = \{1, 2, 4, 5, 9, 10, 11\}$. Then $(1, 2, 4, 5, 9, 10)$ and $(1, 2, 4, 5, 10, 11)$ are both occurrences of $\mathcal{C}$ in A. Any 4-tuple of distinct elements of A is a $\{1, 2, 3, 5\}$-occurrence of $\mathcal{C}$ in A, and $(1, 10, 2, 11)$ is a $\{1, 2, 3, 4\}$-occurrence and a $\{3, 4, 5, 6\}$-occurrence of $\mathcal{C}$ in A.

We are now ready to define the properties of a k-configuration that determine whether we consider them 'good' or 'bad' (as in the outline of our argument given at the beginning of Section 2).

First, we say that a k-configuration is *invalid* if it implies $x_{i_1} - x_{i_2} = 0$ for some $i_1 \neq i_2$ (i.e., it implies two variables are equal), and *valid* otherwise. Note that an invalid k-configuration can never occur in a set A (because of the distinctness condition in the definition of an occurrence). Also note that in a generic solution to a valid k-configuration, all k entries must be distinct.

We say a k-configuration is *AP-containing* if it implies $x_{i_1} - 2x_{i_2} + x_{i_3} = 0$ for some distinct i_1, i_2, i_3 (i.e., it implies three variables form an arithmetic progression), and *AP-free* otherwise.

Most importantly, given any $1 < c \leq 2$, we say a k-configuration $\mathcal{C}$ is *c-heavy* if there exists some $1 \leq m \leq k$ and some m-element subset $I \subseteq \{1, \ldots, k\}$ such that

$$\dim_I(\mathcal{C}) < \frac{(c-1)m + 1}{c},$$

and we call such a subset I a *c-heavy part* of $\mathcal{C}$; we say $\mathcal{C}$ is *c-light* if it is not c-heavy. (This definition was chosen in order to make the expected value calculation in the proof of Claim 2.8 work out.)

Remark 2.2. Note that if $\mathcal{C}$ is c-light, then for any $I \subseteq \{1, \ldots, k\}$ of size $|I| = m$, any collection of linearly independent equations on I implied by $\mathcal{C}$ has size at most

$$m - \dim_I(\mathcal{C}) \leq m - \frac{(c-1)m + 1}{c} = \frac{m - 1}{c}.$$

Equivalently, any collection of t linearly independent equations implied by $\mathcal{C}$ must involve at least $ct + 1$ variables. We will use this observation frequently in our proof of Proposition 1.7 (with $c = 2$).

Finally, we say a k-configuration $\mathcal{C}$ is *c-good* if it is valid, AP-free, and c-light; we say it is *c-bad* otherwise. Note that if a k-configuration $\mathcal{C}$ has any of these three properties (i.e., is valid, AP-free, or c-light), then any collection of difference equalities implied by $\mathcal{C}$ has the same property.

Example 2.3. To illustrate these definitions, we consider a few examples with $c = 2$.

(a) The 4-configuration $\{x_1 - x_2 - x_3 + x_4 = 0,\ x_1 - x_2 + x_3 - x_4 = 0\}$ is invalid, as it implies $x_3 = x_4$. (It is also 2-heavy, with both $\{3, 4\}$ and $\{1, 2, 3, 4\}$ as 2-heavy parts.)

(b) The 5-configuration $\{x_1 - x_2 - x_3 + x_4 = 0,\ x_1 + x_2 - x_3 - x_5 = 0\}$ is valid and 2-light, but it is AP-containing (as it implies $x_5 - 2x_2 + x_4 = 0$) and therefore 2-bad.

(c) The 12-configuration $\{x_1 - x_2 - x_3 + x_4 = 0,\ x_1 - x_2 - x_5 + x_6 = 0,\ x_1 - x_2 - x_7 + x_8 = 0,\ x_1 - x_3 - x_5 + x_7 = 0,\ x_9 - x_{10} - x_{11} + x_{12} = 0\}$ is valid and AP-free, but it is 2-heavy with 2-heavy part $\{1, \ldots, 8\}$ (as it contains 4 linearly independent equations on $\{1, \ldots, 8\}$), and therefore 2-bad.

(d) The 6-configuration $\mathcal{C} = \{x_1 - x_2 - x_5 + x_4 = 0,\ x_1 - x_3 - x_6 + x_4 = 0\}$ — or equivalently, $\{x_1 + x_4 = x_2 + x_5 = x_3 + x_6\}$ — is 2-good, as it is valid, AP-free, and 2-light. The fact that it is 2-light can be seen by pairing x_1 and x_4, x_2 and x_5, and x_3 and x_6 — we can obtain the solutions to $\mathcal{C}$ by freely choosing the values of both variables in one pair and one variable in each of the other pairs (then the values of the remaining two variables are uniquely determined), and for any $I \subseteq \{1, \ldots, 6\}$ of size $m \geq 1$, this gives a way of freely choosing the values of more than half of the variables x_i with indices $i \in I$, thus showing that $\dim_I(\mathcal{C}) \geq \frac{m+1}{2}$.

2.1.3. *Our main steps.* Finally, we state the results of the two main steps of our argument as lemmas and explain why they imply Theorem 1.6.

The following lemma will be the result of the first step of our argument, where we use a random construction to obtain a set A with few distinct differences that 'avoids' all c-bad k-configurations.

Lemma 2.4. *Fix a positive integer k and real numbers $1 < c \le 2$ and $a > 0$. Then for all sufficiently large n (with respect to k, c, and a), there exists an n-element set $A \subseteq \mathbb{R}$ with $|A - A| \le an^c$ such that every k-configuration that occurs in A is c-good.*

The following lemma will be the result of the second step of our argument, in which we show that a generic solution to a c-good k-configuration contains many distinct differences. In fact, we show more generally that a generic solution to any valid k-configuration with high dimension contains many distinct differences.

Lemma 2.5. *Fix a positive integer k, let $\mathcal{C}$ be a valid k-configuration, and let $\dim(\mathcal{C}) = d$. Let $(a_1, \ldots, a_k)$ be a generic solution to $\mathcal{C}$, and let $A' = \{a_1, \ldots, a_k\}$. Then*

$$(6) \qquad |A' - A'| \ge \begin{cases} (d-1)^2 & \text{if } d \le \frac{k}{2} + 1 \\ \binom{k}{2} - (k-d)(k-d+1) & \text{if } d \ge \frac{k}{2} + 1. \end{cases}$$

Note that the two expressions in (6) are equal at $d = \frac{k}{2} + 1$, and both are (weakly) increasing in d; this means that the entire right-hand side of (6) is increasing in d.

We will prove these two lemmas in Subsections 2.2 and 2.3, respectively. For now, we explain why they together imply Theorem 1.6.

Proof of Theorem 1.6. By the comments at the beginning of Section 2, it suffices to show that the sets A given by Lemma 2.4 satisfy the (k, ℓ)-local property for the value of ℓ in Theorem 1.6. To see this, suppose that a set A has the property that every k-configuration occurring in A is c-good. Then for any k-element subset $A' = \{a_1, \ldots, a_k\} \subseteq A$, we can write down the k-configuration $\mathcal{C}$ consisting of *all* the difference equalities satisfied by $(a_1, \ldots, a_k)$. Then $\mathcal{C}$ occurs in A, so it must be c-good; in particular, $\mathcal{C}$ is valid and

$$\dim(\mathcal{C}) \ge \left\lceil \frac{(c-1)k + 1}{c} \right\rceil$$

(by the c-light condition applied to the entire set $\{1, \ldots, k\}$, and the fact that $\dim(\mathcal{C})$ is an integer). Meanwhile $(a_1, \ldots, a_k)$ is a generic solution to $\mathcal{C}$ (as every difference equality it satisfies is in fact *contained* in $\mathcal{C}$), so we can bound $|A' - A'|$ using Lemma 2.5. The bound given by (6) is increasing in d and we have

$$\left\lceil \frac{(c-1)k + 1}{c} \right\rceil \le \left\lceil \frac{k+1}{2} \right\rceil \le \frac{k}{2} + 1,$$

so Lemma 2.5 gives that

$$|A' - A'| \ge \left(\left\lceil \frac{(c-1)k + 1}{c} \right\rceil - 1 \right)^2 = \left\lceil \frac{(c-1)(k-1)}{c} \right\rceil^2 = \ell.$$

This means A indeed satisfies the (k, ℓ)-local property, as desired. $\qquad\qquad\square$

Remark 2.6. For this argument, we do not need the full strength of Lemma 2.4 — the only parts of the c-good condition that we need are that a c-good configuration $\mathcal{C}$ is valid and satisfies $\dim(\mathcal{C}) \ge \frac{(c-1)k+1}{c}$. However, we will need the remaining parts of the c-good condition in our proof of Proposition 1.7 (and we state Lemma 2.4 in this way so that we can use it for that proof as well).

2.2. **Proof of Lemma 2.4.** In this subsection, we prove Lemma 2.4 using a random construction (with the alteration method). Our construction will make use of the following theorem of Behrend.

Theorem 2.7 (Behrend [1]). *For every positive integer n, there exists a set $S \subseteq \{1, \ldots, n\}$ of size $|S| = n^{1-o(1)}$ containing no 3-term arithmetic progression.*

Proof of Lemma 2.4. Assume that n is sufficiently large with respect to k, c, and a (all our asymptotic notation treats k, c, and a as fixed, and the implied constants will depend on them). First, by Theorem 2.7, there exists a set $S \subseteq \{1, \ldots, \lceil an^c \rceil\}$ of size $s = n^{c-o(1)}$ with no 3-term arithmetic progression. We will ultimately take A to be an n-element subset of S, which will automatically imply that $|A - A| \le an^c$ (as we

have $A - A \subseteq S - S \subseteq \{1, \ldots, \lceil an^c \rceil - 1\}$) and that every k-configuration that occurs in A is AP-free. So in order to prove Lemma 2.4, it will suffice to ensure that every k-configuration that occurs in A is c-light. (We automatically have that every k-configuration that occurs in A is valid, as this is true for *any* set.)

Let $p = \frac{3n}{s}$. Note that $p = n^{1-c+o(1)}$, so $p \in (0, 1)$ for all sufficiently large n. Let T be a randomly chosen subset of S where each element is included independently with probability p. Then $\mathbb{E}[|T|] = 3n$, and since $|T|$ is a sum of s independent Bernoulli random variables, by the Chernoff bounds we have $|T| \geq 2n$ with probability $1 - o(1)$.

We will show that with probability $1 - o(1)$, it is possible to destroy all occurrences of c-heavy k-configurations in T by deleting at most n of its elements (which we will then do to obtain our set A). The following claim allows us to do so.

Claim 2.8. *Let $\mathcal{C}$ be a c-heavy k-configuration, and let I be a c-heavy part of $\mathcal{C}$. Then the expected number of I-occurrences of $\mathcal{C}$ in T is $o(n)$.*

Proof. Let $d = \dim_I(\mathcal{C})$. First, we claim that there are at most s^d I-occurrences of $\mathcal{C}$ in S. To see this, without loss of generality assume $I = \{1, \ldots, m\}$ for some m. By solving $\mathcal{C}$ using reduced row echelon form, we can find a list of indices $i_1 < \cdots < i_{d'}$, where $d' = \dim(\mathcal{C})$, such that every choice of values for x_{i_1}, $\ldots$, $x_{i_{d'}}$ in $\mathbb{R}$ corresponds to a unique solution $(x_1, \ldots, x_k)$ to $\mathcal{C}$, and for each $i \in \{1, \ldots, k\}$, the value of x_i in this solution depends on only the chosen values of x_{i_j} for the indices $i_j < i$. Then each choice of values for x_{i_j} over all indices $i_j \leq m$ corresponds to a unique I-solution to $\mathcal{C}$, so there must be exactly $\dim_I(\mathcal{C}) = d$ such indices. Finally, each choice of values for these d variables x_{i_j} in S provides at most one I-solution to $\mathcal{C}$ whose elements are in S, and therefore at most one I-occurrence of $\mathcal{C}$ in S. There are s^d such choices, so there are at most s^d I-occurrences of $\mathcal{C}$ in S.

Now for each I-occurrence of $\mathcal{C}$ in S, the probability that all its elements are included in T is p^m (as it has m distinct elements, each of which is included in T independently with probability p). So by the linearity of expectation, the expected number of I-occurrences of $\mathcal{C}$ in T is at most $s^d \cdot p^m$. Since $s = n^{c-o(1)}$ and $p = n^{1-c+o(1)}$, we have
$$s^d \cdot p^m = n^{cd + m(1-c) + o(1)}.$$
Since $cd + m(1 - c) < 1$ by the definition of a c-heavy part (and this quantity is a constant not depending on n), we have $s^d \cdot p^m = o(n)$, as desired. □

Now for each c-heavy k-configuration $\mathcal{C}$, fix some c-heavy part of $\mathcal{C}$, which we denote by $h(\mathcal{C})$. Then Claim 2.8 implies that for each *fixed* c-heavy $\mathcal{C}$, the number of $h(\mathcal{C})$-occurrences of $\mathcal{C}$ in T has expected value $o(n)$. But the total number of k-configurations is a constant (depending on k but not n), since for any given k, there are only finitely many difference equalities. So the *total* number of $h(\mathcal{C})$-occurrences of $\mathcal{C}$ over all c-heavy $\mathcal{C}$ has expected value $o(n)$ as well. Then by Markov's inequality, this number is at most n with probability $1 - o(1)$.

So with probability $1 - o(1)$, we have $|T| \geq 2n$ and there are at most n total $h(\mathcal{C})$-occurrences of $\mathcal{C}$ in T over all c-heavy k-configurations $\mathcal{C}$; this means that for sufficiently large n, there must exist a set $T \subseteq S$ with both properties. Fix such a set T, and let A be an n-element subset of T obtained by first deleting some element from each $h(\mathcal{C})$-occurrences of each c-heavy $\mathcal{C}$ (which requires us to delete at most n elements), and then deleting additional elements arbitrarily until we are left with exactly n elements.

We have already shown that $|A - A| \leq an^c$ and that no invalid or AP-containing k-configuration can occur in A; this construction guarantees that no c-heavy k-configuration can occur in A (as any occurrence of a c-heavy k-configuration $\mathcal{C}$ would contain a $h(\mathcal{C})$-occurrence of $\mathcal{C}$). So A has all the claimed properties. □

2.3. **Proof of Lemma 2.5.** In this subsection, we prove Lemma 2.5 (our lower bound on the number of distinct differences in a generic solution to a valid k-configuration, in terms of the dimension of the k-configuration). We first establish a few definitions that will be useful in this proof.

Note that equations of the form $x_{i_1} - x_{i_2} = 0$, $2x_{i_1} - 2x_{i_2} = 0$, and $x_{i_1} - 2x_{i_2} + x_{i_3} = 0$ are all considered to be difference equalities, since we do not require the indices in (5) to be distinct. We say a difference equality is *degenerate* if it is of the form $x_{i_1} - x_{i_2} = 0$ for distinct indices i_1 and i_2 (equivalently, in (5) the sets $\{i_1, i_4\}$ and $\{i_2, i_3\}$ are different but not disjoint), and *nondegenerate* otherwise. (Since we assume the k-configuration $\mathcal{C}$ we work with is valid, all difference equalities it implies will be nondegenerate.)

Given a nondegenerate difference equality $(*)$, we say that two variables x_i and x_j appear in $(*)$ *with opposite sign* if $(*)$ can be written as $x_i - x_j = x_{i'} - x_{j'}$ for some $i', j' \in \{1, \ldots, k\}$, and *with the same sign*

if $(*)$ can be written as $x_i + x_j = x_{i'} + x_{j'}$ for some $i', j' \in \{1, \dots, k\}$. This definition matches the intuitive meaning of these terms; in the difference equality (5), x_{i_1} appears with opposite sign as both x_{i_2} and x_{i_3}, and with the same sign as x_{i_4}.

Given distinct indices $i, j \in \{1, \dots, k\}$, we say that a collection of linear equations i-*certifies* j if it implies some nondegenerate difference equality in which x_i and x_j appear with opposite sign. (In particular, we say a single difference equality i-certifies j if x_i and x_j appear in it with opposite sign.)

Our proof of Lemma 2.5 will require the following lemma. The proof of this lemma is somewhat technical, so we defer it to Subsection 2.4 (and first prove Lemma 2.5 assuming this lemma).

Lemma 2.9. *Fix $i \in \{1, \dots, k\}$, and let $\mathcal{S}$ be a valid collection of linearly independent difference equalities that all involve x_i. Then $\mathcal{S}$ i-certifies at most $2|\mathcal{S}|$ indices.*

Proof of Lemma 2.5. We first make a convenient assumption on the ordering of the indices of x_1, ..., x_k: by using reduced row echelon form to solve $\mathcal{C}$ (as a system of equations), we can find d distinct indices i_1, ..., i_d such that every choice of values for x_{i_1}, ..., x_{i_d} gives a unique solution $(x_1, \dots, x_k)$ to $\mathcal{C}$. We assume without loss of generality that these indices are 1, ..., d.

We can calculate $|A' - A'|$ in the following way: for each pair of indices (i, j) with $1 \leq j < i \leq k$, we say (i, j) corresponds to a *repeated difference* if $|a_i - a_j| = |a_{i'} - a_{j'}|$ for some (i', j') with $1 \leq j' < i' \leq k$ that is lexicographically smaller than (i, j) (i.e., either $i' < i$, or $i' = i$ and $j' < j$), and a *new difference* otherwise. Note that if a pair (i, j) corresponds to a repeated difference, then $(a_1, \dots, a_k)$ satisfies a nondegenerate difference equality of the form

$$(7) \qquad\qquad x_i - x_j - x_{i'} + x_{j'} = 0$$

for some $i', j' \leq i$ (here we do not necessarily have $j' < i'$) — i.e., a difference equality on $\{1, \dots, i\}$ in which x_i and x_j occur with opposite sign. Since $(a_1, \dots, a_k)$ is a generic solution to $\mathcal{C}$, this difference equality (7) must be implied by $\mathcal{C}$.

Then $|A' - A'|$ is precisely the number of pairs (i, j) corresponding to new differences. For each index i, we will upper-bound the number of repeated differences contributed by i (i.e., the number of indices $j < i$ for which (i, j) corresponds to a repeated difference) by bounding the number of indices $j < i$ for which $\mathcal{C}$ can imply a difference equality of the form (7). Combining these bounds over all i will then give the desired lower bound on $|A' - A'|$.

We first consider the indices $i \leq d$. Our assumption that every choice of x_1, ..., x_d corresponds to a solution to $\mathcal{C}$ means that $\mathcal{C}$ cannot imply any equation on $\{1, \dots, d\}$, and therefore cannot imply a difference equality of the form (7) for any $j < i \leq d$. So each index $i \leq d$ contributes no repeated differences.

We now consider the indices $i > d$. Fix some index $i > d$, let $\overline{\mathcal{S}_i}$ be the set of all difference equalities on $\{1, \dots, i\}$ involving x_i that are implied by $\mathcal{C}$, and let $\mathcal{S}_i$ be a maximal linearly independent subset of $\overline{\mathcal{S}_i}$. Then $\mathcal{S}_i$ implies every equation in $\overline{\mathcal{S}_i}$, so if a pair (i, j) corresponds to a repeated difference, then $\mathcal{S}_i$ must i-certify j (as the corresponding difference equality (7) must be in $\overline{\mathcal{S}_i}$, and is therefore implied by $\mathcal{S}_i$).

But we have $\dim_{\{1, \dots, i\}}(\mathcal{C}) = d$, since each choice of x_1, ..., x_d corresponds to a unique solution to $\mathcal{C}$. This means we must have $|\mathcal{S}_i| \leq i - d$ (using the fact from linear algebra that a system of linear equations on i variables consisting of s linearly independent equations has a solution space of dimension $i - s$). Then by Lemma 2.9, $\mathcal{S}_i$ i-certifies at most $2(i - d)$ indices, and therefore i contributes at most $2(i - d)$ repeated differences.

We now use these statements to lower-bound $|A' - A'|$. Each index $i \leq d$ contributes no repeated differences, and therefore $i - 1$ new differences; meanwhile, each index $i > d$ contributes at most $2(i - d)$ repeated differences, and therefore at least $i - 1 - 2(i - d) = 2d - i - 1$ new differences. So $|A' - A'|$ — i.e., the total number of pairs corresponding to new differences — is at least

$$\sum_{i=1}^{d}(i-1) + \sum_{i=d+1}^{k} \max(2d - i - 1, 0) = \begin{cases} (d-1)^2 & \text{if } d \leq \frac{k}{2} + 1 \\ \binom{k}{2} - (k-d)(k-d+1) & \text{if } d \geq \frac{k}{2} + 1. \end{cases}$$

(For $d \geq \frac{k}{2} + 1$, it is easier to calculate this bound by noting that there are at most $2(1 + 2 + \dots + (k - d)) = (k-d)(k-d+1)$ pairs corresponding to repeated differences, and therefore at least $\binom{k}{2} - (k-d)(k-d+1)$ pairs corresponding to new differences.) $\qquad\square$

Remark 2.10. Lemma 2.5 is tight for $d \geq \frac{k}{2} + 1$, as equality is achieved by the k-configuration

$$\mathcal{C} = \{x_1 + x_d = x_2 + x_{d+1} = \cdots = x_{k-d+1} + x_k\}.$$

(The fact that this k-configuration is 2-good can be seen by the same argument as in Example 2.3(d).) Each index $1 \leq i \leq d$ contributes no repeated differences, and each index $i > d$ contributes exactly $2(i - d)$ repeated differences, corresponding to the pairs (i, j) and $(i, j - d + 1)$ for all $d \leq j < i$ (as $\mathcal{C}$ implies $x_i + x_{i-d+1} - x_j - x_{j-d+1} = 0$ for all such j, so in any generic solution we have $a_i - a_j = a_{j-d+1} - a_{i-d+1}$ and $a_i - a_{j-d+1} = a_j - a_{i-d+1}$). So there are exactly $(k - d)(k - d + 1)$ pairs corresponding to repeated differences.

2.4. **Proof of Lemma 2.9.** In this subsection, we finally prove Lemma 2.9 (thus completing the proof of Theorem 1.6). Note that if the collection $\mathcal{S}$ in Lemma 2.9 does not imply any difference equalities involving x_i other than the ones already in $\mathcal{S}$, then the conclusion is obvious, as each difference equality in $\mathcal{S}$ individually i-certifies at most 2 indices. The difficulty comes from the fact that $\mathcal{S}$ may imply more difference equalities involving x_i, which may i-certify additional indices; the key observation that allows us to deal with this is the following lemma.

Lemma 2.11. *Let $\mathcal{T} = \{(*_1), \ldots, (*_t)\}$ be a valid collection of t linearly independent difference equalities, and suppose that $\mathcal{T}$ minimally implies a difference equality $(*_{\mathcal{T}})$. Then the following statements hold:*

    (i) *At most $2t + 2$ variables appear in $\mathcal{T}$. Furthermore, if equality holds, then every variable appearing in $\mathcal{T}$ appears in exactly two of $(*_1)$, $\ldots$, $(*_t)$, $(*_{\mathcal{T}})$.*

    (ii) *For every proper subset $\mathcal{T}' \subseteq \mathcal{T}$ of size t', at most $2t'$ variables appear in $\mathcal{T}'$ but not $\mathcal{T} \setminus \mathcal{T}'$.*

Proof. Let $*_1$, $\ldots$, $*_t$, $*_{\mathcal{T}}$ be the contents of $(*_1)$, $\ldots$, $(*_t)$, $(*_{\mathcal{T}})$. Since $\mathcal{T}$ minimally implies $(*_{\mathcal{T}})$, there exist nonzero $\varepsilon_1, \ldots, \varepsilon_t \in \mathbb{R}$ for which

$$(8) \qquad\qquad\qquad * = \varepsilon_1 *_1 + \cdots + \varepsilon_t *_t.$$

We first prove (i). Note that (8) implies that variable that appears in $\mathcal{T}$ must appear in at least two of $(*_1)$, $\ldots$, $(*_t)$, $(*_{\mathcal{T}})$. But each of these $t + 1$ equations involves at most 4 variables, so there are at most $4(t + 1)$ appearances of variables with multiplicity. So the total number of variables appearing in $\mathcal{T}$ is at most $\frac{4(t+1)}{2} = 2t + 2$, and if equality holds then each variable must appear exactly twice.

We now prove (ii). For notational convenience, assume that $\mathcal{T}' = \{(*_1), \ldots, (*_{t'})\}$, and let

$$(9) \qquad\qquad\qquad *_{\mathcal{T} \setminus \mathcal{T}'} = \varepsilon_{t'+1} *_{t'+1} + \cdots + \varepsilon_t *_t.$$

(In words, $*_{\mathcal{T} \setminus \mathcal{T}'}$ is the partial sum of (8) corresponding to only the equations *not* in $\mathcal{T}'$.) Then we have

$$(10) \qquad\qquad\qquad *_{\mathcal{T}} - *_{\mathcal{T} \setminus \mathcal{T}'} = \varepsilon_1 *_1 + \cdots + \varepsilon_t *_t.$$

Let $(*_{\mathcal{T} \setminus \mathcal{T}'})$ be the equation $*_{\mathcal{T} \setminus \mathcal{T}'} = 0$ (this is implied by $\mathcal{T}$, but is not necessarily a difference equality).

We now define three disjoint sets of indices:

- Let I_1 be the set of indices i for which x_i appears in $(*_{\mathcal{T} \setminus \mathcal{T}'})$ but not $(*_{\mathcal{T}})$.
- Let I_2 be the set of indices for which x_i appears in $(*_{\mathcal{T}})$ but not $(*_{\mathcal{T} \setminus \mathcal{T}'})$.
- Let I_3 be the set of indices i for which x_i appears in $\mathcal{T}'$, but not $(*_{\mathcal{T} \setminus \mathcal{T}'})$ or $(*_{\mathcal{T}})$.

Let I_1, I_2, and I_3 have sizes m_1, m_2, and m_3, respectively.

Then for every variable x_i appearing in $\mathcal{T}'$ but not $\mathcal{T} \setminus \mathcal{T}'$, we must have $i \in I_2$ or $i \in I_3$ — if x_i does not appear in $\mathcal{T} \setminus \mathcal{T}'$, then by (9) it certainly does not apply in $(*_{\mathcal{T} \setminus \mathcal{T}'})$, so i is in I_2 if x_i appears in $(*_{\mathcal{T}})$ and I_3 if it does not. This means the number of variables appearing in $\mathcal{T}'$ but not $\mathcal{T} \setminus \mathcal{T}'$ is at most $m_2 + m_3$, so it suffices to show that $m_2 + m_3 \leq 2t'$.

First, we claim that $m_1 + m_2 + 2m_3 \leq 4t'$. To see this, for each $i \in I_1 \cup I_2$, since x_i appears on the left-hand side of (10), it must also appear on the right-hand side; this means x_i appears in at least one equation in $\mathcal{T}'$. Meanwhile, for each $i \in I_3$, since x_i does not appear on the left-hand side of (10) but does appear in at least one summand on the right-hand side, it must appear in at least two; this means x_i appears in at least *two* equations in $\mathcal{T}'$. Since each of the t' equations in $\mathcal{T}'$ involves at most 4 variables, this means that $m_1 + m_2 + 2m_3 \leq 4t'$, as claimed.

On the other hand, we will show that $m_1 \geq m_2 - 1$. First, $(*_{\mathcal{T}})$ involves at most 4 variables, as it is a difference equality. Meanwhile, we claim $(*_{\mathcal{T} \setminus \mathcal{T}'})$ involves at least 3 variables. To see this, first note that $*_{\mathcal{T} \setminus \mathcal{T}'}$ is not simply 0 (as $\mathcal{T} \setminus \mathcal{T}'$ is nonempty and its equations are linearly independent). However,

its coefficients must sum to 0, as this is true of every summand in (8). So $(*_{\mathcal{T}\setminus\mathcal{T}'})$ cannot involve only 1 variable, and if it involved only 2 variables, then it would be of the form $\varepsilon x_i - \varepsilon x_j = 0$ for some $\varepsilon \neq 0$ and $i \neq j$, contradicting the validity of $\mathcal{T}$. So $(*_{\mathcal{T}\setminus\mathcal{T}'})$ must involve at least 3 variables. Finally, $m_1 - m_2$ is the difference between the numbers of variables involved in $(*_{\mathcal{T}\setminus\mathcal{T}'})$ and $(*_{\mathcal{T}})$, so $m_1 - m_2 \geq 3 - 4 = -1$.

Combining these two bounds, we have $2(m_2 + m_3) - 1 \leq m_1 + m_2 + 2m_3 \leq 4t'$. Since $2(m_2 + m_3)$ and $4t'$ are both even, this means $2(m_2 + m_3) \leq 4t'$, so $m_2 + m_3 \leq 2t'$, as desired. $\square$

We are now ready to prove Lemma 2.9.

Proof of Lemma 2.9. Write down all difference equalities in $\mathcal{S}$. For each subset of $\mathcal{S}$ that minimally implies some difference equality also involving x_i, draw a box around this subset. We say the *size* of a box is the number of equations it contains, and we say a box is *big* if it has size greater than 1 (each equation in $\mathcal{S}$ is automatically in a box of size 1, as it minimally implies itself). An example is given in Figure 2.
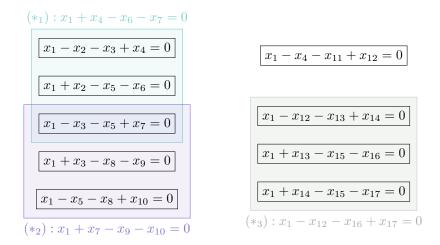

FIGURE 2. This figure shows an example of a possible $\mathcal{S}$ and the corresponding boxes. Here $i = 1$, and $\mathcal{S}$ consists of the equations written in black. Each equation of $\mathcal{S}$ is in a box of size 1, and we have three boxes of size 3, which are drawn in blue, purple, and green and labelled with the difference equalities that they minimally imply (named $(*_1)$, $(*_2)$, and $(*_3)$, respectively). For example, the blue box minimally implies $(*_1)$ because $(x_1 - x_2 - x_3 + x_4) + (x_1 + x_2 - x_5 - x_6) - (x_1 - x_3 - x_5 + x_7) = x_1 + x_4 - x_6 - x_7$.

We will work with the 'connected components' of this picture, which we refer to as 'blobs' — to formalize this, consider the graph on $\mathcal{S}$ in which two equations of $\mathcal{S}$ are adjacent if and only if there is some box containing both of them. We say a *blob* is a connected component of this graph, and the *size* of a blob is again the number of equations it contains. In the example in Figure 2, we have one blob of size 5 (the union of the blue and purple boxes), one blob of size 3 (the green box), and one blob of size 1 (the one equation not in any big box).

Then the blobs form a partition of $\mathcal{S}$ such that if $\mathcal{S}$ i-certifies an index j, then in fact some blob must i-certify j (every difference equality involving x_i that is implied by $\mathcal{S}$ is by definition implied by some box, and therefore by the blob containing this box). We will show that for each blob, the number of indices it i-certifies is at most twice its size.

Clearly any blob of size 1 can i-certify at most 2 indices. Meanwhile, we will show that any blob of size $t > 1$ can only *involve* at most $2t$ variables other than x_i, and therefore can i-certify at most $2t$ indices.

Any blob $\mathcal{T}$ of size $t > 1$ can be written as a union of big boxes $\mathcal{T}_0, \ldots, \mathcal{T}_r$ such that for each $1 \leq p \leq r$, the box $\mathcal{T}_p$ has nonempty intersection with $\mathcal{T}_0 \cup \cdots \cup \mathcal{T}_{p-1}$. Let $t_0 = |\mathcal{T}_0|$, and for each $1 \leq p \leq r$, let $t_p = |\mathcal{T}_p \setminus (\mathcal{T}_0 \cup \cdots \cup \mathcal{T}_{p-1})|$; then $t = t_0 + \cdots + t_r$. (Intuitively, we can imagine 'building' $\mathcal{T}$ by starting with some big box $\mathcal{T}_0$, and then repeatedly adding in big boxes $\mathcal{T}_1, \ldots, \mathcal{T}_r$ such that at each step, the next big box we add overlaps with some previously added box; then for each p, we define t_p as the number of new equations we add to the growing blob when we add in $\mathcal{T}_p$. In the example in Figure 2, for the blob of size 5, we can take $\mathcal{T}_0$ to be the blue box and $\mathcal{T}_1$ to be the purple box; then $t_0 = 3$ and $t_1 = 2$.)

First, Lemma 2.11(i) implies that $\mathcal{T}_0$ contains at most $2t_0 + 2$ variables. Furthermore, equality cannot hold, as x_i appears in all $t_0 + 1 \geq 3$ of the relevant equations, so in fact $\mathcal{T}_0$ contains at most $2t_0 + 1$ variables.

Meanwhile, for each $1 \leq p \leq r$, Lemma 2.11(ii) applied to $\mathcal{T}_p \setminus (\mathcal{T}_0 \cup \cdots \cup \mathcal{T}_{p-1})$ (which is a proper subset of $\mathcal{T}_p$) gives that there are at most $2t_p$ variables that appear in $\mathcal{T}_p$ but not $\mathcal{T}_0 \cup \cdots \cup \mathcal{T}_{p-1}$. (Rephrased in terms of the process of building $\mathcal{T}$ one box at a time, this states that when we add in the big box $\mathcal{T}_p$ — thus adding in t_p new equations to our growing blob — these t_p new equations give us at most $2t_p$ new variables.)

So the total number of variables that appear in $\mathcal{T}$ is at most

$$(2t_0 + 1) + 2t_1 + \cdots + 2t_p = 2t + 1.$$

Since one of these variables is x_i, this means at most $2t$ variables other than x_i appear in $\mathcal{T}$, as desired.

So for every blob, the number of indices it i-certifies is at most twice its size. Since the blobs form a partition of $\mathcal{S}$, this means the *total* number of indices i-certified by any blob is at most $2|\mathcal{S}|$, as desired. $\qquad\square$

Thus we have shown Lemma 2.9, which completes the proof of Theorem 1.6 (our family of upper bounds on $g(n, k, \ell)$).

3. Proof of Proposition 1.7 (stronger quadratic threshold for odd k)

In this section, we prove Proposition 1.7 (our improvement on the value of ℓ in Theorem 1.6 when $c = 2$ and k is odd). We use the same random construction as in our proof of Theorem 1.6 (i.e., we use Lemma 2.4 as written), but improve our analysis of the number of distinct differences in a generic solution to a 2-good k-configuration when k is odd — we replace Lemma 2.5 with the following bound.

Lemma 3.1. *Fix an odd positive integer k, and let $\mathcal{C}$ be a 2-good k-configuration. Let $(a_1, \ldots, a_k)$ be a generic solution to $\mathcal{C}$, and let $A' = \{a_1, \ldots, a_k\}$. Then*

$$|A' - A'| \geq \frac{(k+1)^2}{4} - 4.$$

For comparison, Lemma 2.5 gives only that $|A' - A'| \geq \frac{(k-1)^2}{4}$.

Together, Lemmas 2.4 and 3.1 immediately imply Proposition 1.7 for the same reason that Lemmas 2.4 and 2.5 imply Theorem 1.6 — Lemma 3.1 implies that the sets A given by Lemma 2.4 (when $c = 2$ and k is odd) satisfy the (k, ℓ)-local property for the value of ℓ in Proposition 1.7.

Similarly to our proof of Lemma 2.5, our proof of Lemma 3.1 will rely on two lemmas whose proofs are rather technical. In Subsection 3.1, we state these lemmas and prove Lemma 3.1 assuming them. In Subsection 3.2, we prove a few observations that will be useful in the proofs of both lemmas. Finally, in Subsection 3.3 we prove the first technical lemma, and in Subsection 3.4 we prove the second.

3.1. **Proof of Lemma 3.1.** Our proof of Lemma 3.1 will make use of the following definition (in addition to the ones given at the beginning of Subsection 2.3): we say that two distinct nondegenerate difference equalities $(*_1)$ and $(*_2)$ are *difference-aligned at* (i, j) for distinct indices $i, j \in \{1, \ldots, k\}$ if the variables x_i and x_j appear with opposite sign in both $(*_1)$ and $(*_2)$, and *sum-aligned at* (i, j) if x_i and x_j appear with the same sign in both $(*_1)$ and $(*_2)$. For example, $x_1 - x_2 - x_3 + x_4 = 0$ and $x_1 - x_2 - x_5 + x_6 = 0$ are difference-aligned at $(1, 2)$, while $x_1 - x_2 - x_3 + x_4 = 0$ and $x_1 + x_4 - x_5 - x_6 = 0$ are sum-aligned at $(1, 4)$. We say $(*_1)$ and $(*_2)$ are *difference-aligned* or *sum-aligned* if there exist indices i and j for which they are difference-aligned or sum-aligned at (i, j).

Note that if a k-configuration $\mathcal{C}$ is valid and AP-free, then all difference equalities it implies must involve exactly 4 distinct variables (as a difference equality *not* involving 4 distinct variables is of one of the forms $x_{i_1} - x_{i_2} = 0$, $2x_{i_1} - 2x_{i_2} = 0$, or $x_{i_1} - 2x_{i_2} + x_{i_3} = 0$; the first two contradict the fact that $\mathcal{C}$ is valid, and the third contradicts the fact that $\mathcal{C}$ is AP-free.)

We first make a simple observation that will make these definitions more convenient to work with. (This observation will also be useful in several other ways for the proofs of Lemma 3.3 and 3.4.)

Lemma 3.2. *Let $(*_1)$ and $(*_2)$ be distinct difference equalities such that $\{(*_1), (*_2)\}$ is valid and AP-free. Then $(*_1)$ and $(*_2)$ share at most 2 variables.*

Proof. Assume for contradiction that $(*_1)$ and $(*_2)$ share at least 3 variables, so in total they involve at most 5 variables. Let $*_1$ and $*_2$ be the contents of $(*_1)$ and $(*_2)$. We can assume without loss of generality that at least 2 of the shared variables have the same coefficient in both $*_1$ and $*_2$ (either at least 2 have the same

coefficient, or at least 2 have opposite coefficients, and in the latter case we can replace $*_2$ by its negative to reach the former case).

Now let $* = *_1 - *_2$, and let $(*)$ be the equation $* = 0$ (which is implied by $\{(*_1), (*_2)\}$). The variables that appear with the same coefficient in $*_1$ and $*_2$ must cancel out of $*$, so $(*)$ involves at most $5 - 2 = 3$ variables. Furthermore, $*$ is not identically 0 (as $(*_1)$ and $(*_2)$ are distinct), but its coefficients sum to 0 and are all ± 1 or ± 2. So $(*)$ must be of the form $x_{i_1} - x_{i_2} = 0$, $2x_{i_1} - 2x_{i_2} = 0$, or $x_{i_1} - 2x_{i_2} + x_{i_3} = 0$, contradicting the fact that $\{(*_1), (*_2)\}$ is valid and AP-free.   □

In particular, Lemma 3.2 means that if $(*_1)$ and $(*_2)$ share some variable x_i and are difference-aligned or sum-aligned, then they must be difference-aligned or sum-aligned at (i, j) for some index $j \neq i$. In our argument, we will frequently work with collections of difference equalities that all involve some fixed variable x_i; this observation will allow us to simply write that two equations in such a collection are difference-aligned or sum-aligned when we mean that they are difference-aligned or sum-aligned at (i, j) for some j.

We now state the two technical lemmas we need for our proof of Lemma 3.1. The first lemma gives a characterization of the 2-good collections $\mathcal{S}$ that achieve equality or near-equality in Lemma 2.9. (As stated, the lemma only gives a necessary condition on $\mathcal{S}$; it is not hard to show that the converse of the lemma is true as well — i.e., this condition is sufficient — but we do not need this for our purposes.)

Lemma 3.3. *Fix $i \in \{1, \ldots, k\}$, and let $\mathcal{S}$ be a 2-good collection consisting of s linearly independent difference equalities that all involve x_i, such that $\mathcal{S}$ i-certifies at least $2s - 1$ indices. Then we can find some $\mathcal{S}'$, also consisting of s linearly independent difference equalities that all involve x_i, such that $\mathcal{S}'$ is equivalent to $\mathcal{S}$ and the following statements hold:*

(i) If $\mathcal{S}$ i-certifies exactly $2s$ indices, then no two equations in $\mathcal{S}'$ are difference-aligned.
(ii) If $\mathcal{S}$ i-certifies exactly $2s - 1$ indices, then exactly one pair of equations in $\mathcal{S}'$ is difference-aligned.

The second technical lemma essentially considers a collection of difference equalities of the form produced by Lemma 3.3, and bounds the number of difference equalities *not* involving x_i that such a collection implies.

Lemma 3.4. *Fix $i \in \{1, \ldots, k\}$, and suppose that $\mathcal{S}$ is a 2-good collection of s linearly independent difference equalities that all involve x_i, such that at most one pair of equations in $\mathcal{S}$ is difference-aligned.*

(i) If no two equations in $\mathcal{S}$ are difference-aligned and not all are sum-aligned, then $\mathcal{S}$ implies at most $\binom{s-1}{2}$ difference equalities that do not involve x_i.
(ii) If exactly one pair of equations in $\mathcal{S}$ is difference-aligned, then $\mathcal{S}$ implies at most $\binom{s-1}{2} + 1$ difference equalities that do not involve x_i.
(iii) If all equations in $\mathcal{S}$ are sum-aligned, then $\mathcal{S}$ implies at most $\binom{s}{2}$ difference equalities that do not involve x_i.

Note that the cases (ii) and (iii) are disjoint (and of course (i) is also disjoint from both of them) — by Lemma 3.2, $\mathcal{S}$ cannot contain two equations that are simultaneously difference-aligned and sum-aligned. In fact, equality always holds in (iii), but we do not need this for our proof.

We now use these lemmas to prove Lemma 3.1.

Proof of Lemma 3.1. Let $\dim(\mathcal{C}) = d$; since $\mathcal{C}$ is 2-good, we must have $d \geq \frac{k+1}{2}$. First note that if $d \geq \frac{k+3}{2}$, then we are done by Lemma 2.5, as we have

$$|A' - A'| \geq \binom{k}{2} - \frac{(k-3)(k-1)}{4} = \frac{(k+1)^2}{4} - 1 > \frac{(k+1)^2}{4} - 4.$$

So we now assume that $d = \frac{k+1}{2}$.

In order to lower-bound $|A' - A'|$, we use the same notion of new and repeated differences as in our proof of Lemma 2.5 — we will show that the number of pairs corresponding to repeated differences is at most

$$(k-d)(k-d-1) + 3 = \frac{(k-3)(k-1)}{4} + 3,$$

and therefore the number of pairs corresponding to new differences is at least

$$\binom{k}{2} - \frac{(k-3)(k-1)}{4} - 3 = \frac{(k+1)^2}{4} - 4.$$

(For comparison, in Lemma 2.5 we proved the number of pairs corresponding to repeated differences was at most $(k-d)(k-d+1)$; so we wish to improve this bound by $2(k-d)-3$.) As before, if a pair (i,j) with $i > j$ corresponds to a repeated difference, then $\mathcal{C}$ must imply some nondegenerate difference equality

$$x_i - x_j - x_{i'} + x_{j'} = 0 \tag{11}$$

with $i', j' < i$, i.e., a nondegenerate difference equality on $\{1, \ldots, i\}$ in which x_i and x_j appear with opposite sign. Similarly to in the proof of Lemma 2.5, we will first make a certain assumption (without loss of generality) on the ordering of the indices of $x_1, \ldots, x_k$ (although this assumption will be different from the one we used to prove Lemma 2.5), and then bound the number of pairs (i,j) for which $\mathcal{C}$ implies a difference equality of the form (11).

Our assumption on the ordering of indices will make use of the following definition. Given some set $I \subseteq \{1, \ldots, k\}$ and index $i \in I$, we say that I is *i-saturated* if it satisfies both of the following conditions:

- We have $\dim_I(\mathcal{C}) = d$.
- There are at least $2(|I| - d) - 1$ indices $j \in I$ for which $\mathcal{C}$ implies a difference equality on I in which x_i and x_j appear with opposite sign.

We say that I is *saturated* if it is i-saturated for some $i \in I$.

Note that there must exist some saturated set of indices — by again using reduced row echelon form to solve $\mathcal{C}$, we can find d distinct indices $i_1, \ldots, i_d$ such that every choice of $x_{i_1}, \ldots, x_{i_d}$ corresponds to a unique solution $(x_1, \ldots, x_k)$ to $\mathcal{C}$, and then the set $\{i_1, \ldots, i_d\}$ is saturated (as the second condition is trivially satisfied). So we can let $I \subseteq \{1, \ldots, k\}$ be a *maximal* saturated set (meaning that I is saturated, but no proper superset of I is). Without loss of generality, assume that $I = \{1, \ldots, i_0\}$ for some $d \le i_0 \le k$, and that I is i_0-saturated — this is the assumption we make on the ordering of the variables.

Then the structure of our argument is as follows:

(1) We will individually upper-bound the number of repeated differences contributed by each $i > i_0$ using the fact that $\{1, \ldots, i\}$ is not i-saturated. (This bound will be 2 less than the bound we would get using the same argument as in our proof of Lemma 2.5.)

(2) We will bound the number of repeated differences contributed by $i = i_0$ in the same way as in our proof of Lemma 2.5.

(3) Finally, we will bound the *total* number of repeated differences contributed by all $i < i_0$ by combining the fact that $I = \{1, \ldots, i_0\}$ is i_0-saturated together with Lemmas 3.3 and 3.4. The bound we get from this will 'usually' be significantly less than the bound we would get using a similar argument to the one in our proof of Lemma 2.5; then combining it with (1) and (2) gives the desired bound.

(4) The one case in which this does *not* happen (i.e., the bound in step (3) is not good enough) is when we are in case (iii) when we apply Lemma 3.4. When this happens, roughly speaking, the difference equalities $\mathcal{C}$ implies on $\{1, \ldots, i_0\}$ have the same structure as in the equality case of Lemma 2.5 described in Remark 2.10, so it is not possible to get a better bound on the number of repeated differences contributed by $i \le i_0$. However, using this structure, we will show that the index $i_0 + 1$ exists and contributes much fewer repeated differences than our original bound in (1); this adjustment to (1) then gives the desired bound.

We now proceed with the argument. For step (1), for each $i > i_0$, the fact that $\{1, \ldots, i\}$ is not i-saturated means that there are strictly fewer than $2(i - d) - 1$ indices for which $\mathcal{C}$ implies a difference equality of the form (11), and therefore i contributes at most $2(i - d - 1)$ repeated differences.

Now, to perform steps (2) and (3), let $\overline{\mathcal{S}}$ be the set of difference equalities on $I = \{1, \ldots, i_0\}$ involving x_{i_0} that are implied by $\mathcal{C}$, and let $\mathcal{S}$ be a maximal linearly independent subset of $\overline{\mathcal{S}}$. As in the proof of Lemma 2.5, since $\dim_I(\mathcal{C}) = d$, we must have $|\mathcal{S}| \le i_0 - d$. On the other hand, since I is i_0-saturated, $\mathcal{S}$ must i_0-certify at least $2(i_0 - d) - 1$ indices, so by Lemma 2.9 we must have $|\mathcal{S}| \ge i_0 - d$. So we must have $|\mathcal{S}| = i_0 - d$, and $\mathcal{S}$ must i_0-certify at least $2|\mathcal{S}| - 1$ indices. Then we can apply Lemma 3.3 to $\mathcal{S}$; let $\mathcal{S}'$ be the collection of difference equalities that it provides, so that at most one pair of equations in $\mathcal{S}'$ is difference-aligned. Since $\mathcal{S}$ and $\mathcal{S}'$ are equivalent (and each is a collection of linearly independent equations), we have $|\mathcal{S}'| = |\mathcal{S}| = i_0 - d$.

For step (2), the number of repeated differences contributed by the index $i = i_0$ is at most the number of indices i_0-certified by $\mathcal{S}$, which is either $2(i_0 - d)$ or $2(i_0 - d) - 1$.

Meanwhile, for step (3), the total number of pairs (i, j) with $i < i_0$ corresponding to repeated differences is at most twice the number of difference equalities on $\{1, \ldots, i_0 - 1\}$ implied by $\mathcal{C}$ — the equation (11) corresponding to each such pair is a difference equality on $\{1, \ldots, i_0 - 1\}$ implied by $\mathcal{C}$, and each such difference equality corresponds to at most 2 such pairs (i, j) (as i must be the largest index present in the difference equality, and j must be one of the two indices for which x_i and x_j appear with opposite sign).

But as stated earlier, since $\dim_I(\mathcal{C}) = d$, any linearly independent collection of equations on I implied by $\mathcal{C}$ has size at most $i_0 - d$. Since $\mathcal{S}'$ already has size $i_0 - d$, this means every equation on I implied by $\mathcal{C}$ is in fact implied by $\mathcal{S}'$. So the number of difference equalities on $\{1, \ldots, i_0 - 1\}$ implied by $\mathcal{C}$ is precisely the number of difference equalities implied by $\mathcal{S}'$ not involving x_{i_0}, which we can upper-bound using Lemma 3.4 (since at most one pair of equations in $\mathcal{S}'$ is difference-aligned).

We now consider the following three cases (based on which case of Lemma 3.4 applies to $\mathcal{S}'$).

(a) First suppose that no equations in $\mathcal{S}'$ are difference-aligned and not all are sum-aligned. Then $\mathcal{S}$ must fall under the first case of Lemma 3.3 (since $\mathcal{S}'$ has no difference-aligned pairs if $\mathcal{S}$ falls under the first case, and one difference-aligned pair if $\mathcal{S}$ falls under the second), so it i_0-certifies exactly $2(i_0 - d)$ indices. Meanwhile, by Lemma 3.4(i), $\mathcal{S}'$ implies at most $\binom{i_0 - d - 1}{2}$ difference equalities not involving x_{i_0}.

Now combining our bounds from the steps (1), (2), and (3), we have that the total number of pairs (i, j) corresponding to repeated differences (over all indices $1 \leq i \leq k$) is at most

$$2\binom{i_0 - d - 1}{2} + 2(i_0 - d) + \sum_{i=i_0+1}^{k} 2(i - d - 1) = (k - d)(k - d - 1) + 2.$$

(One intuitive way to see why this calculation works is to compare this to the proof of Lemma 2.5. There, we used the bound that each index $i > d$ contributes at most $2(i - d)$ repeated differences, giving a bound of $2(1 + 2 + \cdots + (k - d)) = (k - d)(k - d + 1)$ on the total number of repeated differences. Here we save 2 for each index $i > i_0$; and for the indices $i < i_0$ we have a *total* contribution of $2\binom{i_0 - d - 1}{2}$ instead of $2(1 + 2 + \cdots + (i_0 - d - 1)) = 2\binom{i_0 - d}{2}$, which means we save $2(i_0 - d - 1)$. So in total, we save $2(k - i_0) + 2(i_0 - d - 1) = 2(k - d) - 2$ compared to the bound in Lemma 2.5.)

(b) Now suppose that exactly one pair of equations in $\mathcal{S}'$ is difference-aligned. Then $\mathcal{S}$ must fall under the second case of Lemma 3.3, so it i_0-certifies exactly $2(i_0 - d) - 1$ indices. Meanwhile, by Lemma 3.4(ii), $\mathcal{S}'$ implies at most $\binom{i_0 - d - 1}{2} + 1$ difference equalities not involving x_{i_0}. Again combining our three bounds, the total number of pairs corresponding to repeated differences is at most

$$2\binom{i_0 - d - 1}{2} + 2 + (2(i_0 - d) - 1) + \sum_{i=i_0+1}^{k} 2(i - d - 1) = (k - d)(k - d - 1) + 3.$$

(c) Finally, suppose that all equations in $\mathcal{S}'$ are sum-aligned. This means none are difference-aligned, so $\mathcal{S}$ again falls under the first case of Lemma 3.3, and therefore i_0-certifies exactly $2(i_0 - d)$ indices. Meanwhile, by Lemma 3.4(iii), $\mathcal{S}'$ implies at most $\binom{i_0 - d}{2}$ difference equalities not involving x_{i_0}.

We now perform step (4). Since $\mathcal{S}'$ is valid and AP-free, and all its equations are sum-aligned, there must exist some $j_0 < i_0$ for which all its equations in $\mathcal{S}'$ are of the form

(12) $$x_{i_0} + x_{j_0} = x_i + x_j$$

for distinct $i, j < i_0$, and the $i_0 - d$ pairs $\{i, j\}$ over all these equations must be disjoint from each other and from $\{i_0, j_0\}$.

In particular, $\mathcal{S}'$ involves exactly $2(i_0 - d + 1)$ indices, all of which are at most i_0. This means we must have $2(i_0 - d + 1) \leq i_0$, so $i_0 \leq 2d - 2 = k - 1$. So $i_0 + 1$ is indeed a valid index.

We now show that the index $i = i_0 + 1$ contributes at most 2 repeated differences. (This is a significant improvement on the bound of $2(i_0 - d)$ from step (1) used in the previous two cases, and will make up for the fact that our bound from step (3) is significantly larger in this case.)

Let $I' = \{1, \ldots, i_0 + 1\}$. If the index $i_0 + 1$ contributes no repeated differences, then there is nothing to prove; now assume that it contributes at least one repeated difference, so there exist

distinct $i_1, i_2, i_3, i_4 \le i_0$ such that $\mathcal{C}$ implies

(13) $$x_{i_0+1} - x_{i_1} - x_{i_2} + x_{i_3} = 0.$$

Claim 3.5. *Neither x_{i_1} nor x_{i_2} appears in $\mathcal{S}'$.*

Proof. We will show that if x_{i_1} appears in $\mathcal{S}'$, then I' is i_1-saturated, contradicting the fact that I is a *maximal* saturated set. (Then by symmetry x_{i_2} cannot appear in $\mathcal{S}'$ either.)

Assume for contradiction that x_{i_1} appears in $\mathcal{S}'$, and let

$$x_{i_0} + x_{j_0} = x_{i_1} + x_{j_1}$$

be the equation it appears in. Then $\mathcal{S}'$ also implies the difference equalities

$$x_{i_1} + x_{j_1} = x_i + x_j$$

for each of the $i_0 - d - 1$ pairs $\{i, j\}$ in (12) except $\{i_1, j_1\}$. Then since x_{i_0+1} appears with opposite sign as x_{i_1} in (13), we have at least $2(i_0 - d) + 1 = 2(i_0 - d + 1) - 1$ variables that occur with opposite sign as x_{i_1} in some difference equality on I' implied by $\mathcal{C}$ — namely, all variables appearing in $\mathcal{S}'$ other than x_{i_1} itself and x_{j_1}, together with x_{i_0+1}. So I' is i_1-saturated, which contradicts the maximality of I. $\qquad\square$

Then we claim that $\mathcal{C}$ cannot imply any difference equality on I' involving x_{i_0+1} other than (13). To see this, assume that $\mathcal{C}$ implies such a difference equality $(*)$. First, $(*)$ must be implied by $\mathcal{S}'$ together with (13), since $\mathcal{S}'$ and (13) together form a collection of $i_0 - d + 1$ linearly independent equations on I' (and $\dim_{I'}(\mathcal{C}) = d$, so *any* collection of linearly independent equations on I' implied by $\mathcal{C}$ has size at most $i_0 - d + 1$). But the linear combination used to obtain $(*)$ must use (13) in order for the variable x_{i_0+1} to appear; then since x_{i_1} and x_{i_2} appear in (13) but not in $\mathcal{S}'$, they must appear in $(*)$. This means (13) and $(*)$ share the three variables x_{i_0+1}, x_{i_1}, and x_{i_2}, contradicting Lemma 3.2.

So the index $i_0 + 1$ only contributes 2 repeated differences (corresponding to the pairs $(i_0 + 1, i_1)$ and $(i_0 + 1, i_2)$), as desired.

Then combining steps (1), (2), (3), and (4), the total number of pairs corresponding to repeated differences is at most

$$2\binom{i_0 - d}{2} + 2(i_0 - d) + 2 + \sum_{i=i_0+2}^{k} 2(i - d - 1) = (k - d)(k - d - 1) + 2.$$

(This calculation can again be understood by a comparison to Lemma 2.5 — here we have the same bounds for the total contribution of all indices $i \le i_0$, we save $2(i_0 - d)$ in the bound for the contribution of $i_0 + 1$, and we save 2 for each index $i > i_0 + 1$, for a total saving of $2(i_0 - d) + 2(k - i_0 - 1) = 2(k - d) - 2$.)

So in all cases, at most $(k - d)(k - d - 1) + 3$ pairs correspond to repeated differences, as desired. $\qquad\square$

Remark 3.6. Lemma 3.1 is tight for $k \ge 7$ — letting $k = 2d - 1$, the k-configuration

$$\mathcal{C} = \{x_1 + x_d = x_2 + x_{d+1} = \cdots = x_{d-1} + x_{2d-2}, \, x_1 - x_2 - x_3 + x_{2d-1} = 0\}$$

achieves equality. The pairs corresponding to repeated differences are precisely (i, j) and $(i, j - d + 1)$ for $d \le j < i \le 2d - 2$, as well as $(2d - 1, 2)$, $(2d - 1, 3)$, and $(2d - 1, d)$ — this last pair corresponds to a repeated difference because the equations $x_1 + x_d - x_2 - x_{d+1} = 0$ and $x_1 - x_2 - x_3 + x_{2d-1} = 0$ together imply $x_3 + x_d - x_{d+1} - x_{2d+1} = 0$. So there are exactly

$$2\binom{d-1}{2} + 3 = \frac{(k-1)(k-3)}{4} + 3$$

repeated differences. (Note that this k-configuration does not satisfy the same assumption on ordering that we make in the proof of Lemma 3.1 — here the maximal saturated set is $\{1, \ldots, 2d-1\}$, which is 2-saturated and 3-saturated (and d-saturated) — so we would have to reorder the indices in order to run that proof.)

3.2. **Some useful observations.** The proofs of Lemmas 3.3 and 3.4 will both involve analyzing subsets of the 2-good collection $\mathcal{S}$ that minimally imply certain other difference equalities; the following two lemmas will be useful for this analysis.

Lemma 3.7. *Let $\mathcal{T} = \{(*_1), \ldots, (*_t)\}$ be a 2-good collection of t linearly independent difference equalities, and suppose that $\mathcal{T}$ minimally implies some difference equality $(*_\mathcal{T})$. Then the following two statements hold:*

(i) *Either $\mathcal{T}$ involves exactly $2t + 2$ variables and each of these variables appears exactly twice among $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$; or $\mathcal{T}$ contains exactly $2t + 1$ variables, one appears exactly four times, and the others all appear exactly twice.*

(ii) *Let $*_1$, ..., $*_t$, $*_\mathcal{T}$ be the contents of $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$. Then there exist $\varepsilon_1, \ldots, \varepsilon_t \in \{\pm 1\}$ such that we have*
$$*_\mathcal{T} = \varepsilon_1 *_1 + \cdots + \varepsilon_t *_t.$$

Proof. We will first prove a weaker version of (i) (with one additional possibility); then using this weaker version of (i), we will prove (ii); finally, we will use (ii) to eliminate the additional possibility and prove (i).

Let m be the number of variables appearing in $\mathcal{T}$, and let $e_1, \ldots, e_m$ be the numbers of equations among $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$ that each variable appears in.

On one hand, $\mathcal{T}$ consists of t linearly independent equations on m variables; so since $\mathcal{T}$ is 2-light, by Remark 2.2 we must have $m \geq 2t + 1$.

On the other hand, we can use the same idea as in the proof of Lemma 2.11 — first, each equation involves exactly 4 variables (and every variable appearing in $(*_\mathcal{T})$ also appears in $\mathcal{T}$), so
$$e_1 + \cdots + e_m = 4(t + 1).$$
Meanwhile, since $\mathcal{T}$ minimally implies $(*_\mathcal{T})$, there must exist nonzero $\varepsilon_1, \ldots, \varepsilon_t \in \mathbb{R}$ such that

(14) $$*_\mathcal{T} = \varepsilon_1 *_1 + \cdots + \varepsilon_t *_t,$$

which means each of e_1, ..., e_m must be at least 2 — any variable appearing in only one of $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$ would not cancel out of (14). This immediately gives $m \leq 2t + 2$, so m must be $2t + 1$ or $2t + 2$. Furthermore, if $m = 2t + 2$, then e_1, ..., e_m must all be 2; if $m = 2t + 1$, then either one of e_1, ..., e_m is 4 and the rest are 2, or two are 3 and the rest are 2. In order to prove (i), it suffices to eliminate the last possibility; we will do so after proving (ii).

We now prove (ii), i.e., that in (14) the coefficients ε_1, ..., ε_t must all be ± 1. Note that we have already shown that all but at most 2 variables appearing among $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$ appear exactly twice; we will use this in our proof of (ii).

Consider the graph $\mathcal{G}$ with vertex set $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$, where two equations are adjacent if some variable appears in both of these equations and no others. If two equations are adjacent in $\mathcal{G}$, then their coefficients in (14) must have equal magnitude — otherwise the shared variable would not cancel out of (14). So in order to prove (ii), it suffices to show that $\mathcal{G}$ is connected.

Assume for contradiction that $\mathcal{G}$ is not connected. Then it must have some connected component not containing $(*_\mathcal{T})$; let $\mathcal{T}' \subseteq \mathcal{T}$ be the set of equations this connected component consists of. For notational convenience, assume that $\mathcal{T}' = \{(*_1), \ldots, (*_{t'})\}$ for some $t' \leq t$; then let

(15) $$*_{\mathcal{T}'} = \varepsilon_1 *_1 + \cdots + \varepsilon_{t'} *_{t'}.$$

(In words, $*_{\mathcal{T}'}$ is the partial sum of (14) corresponding to only the equations in $\mathcal{T}'$.) Then we also have

(16) $$*_{\mathcal{T}'} = *_\mathcal{T} - \varepsilon_{t'+1} *_{t'+1} - \cdots - \varepsilon_t *_t.$$

Let $(*_{\mathcal{T}'})$ be the equation $*_{\mathcal{T}'} = 0$ (which is implied by $\mathcal{T}$).

We claim that no variable that appears in exactly two of $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$ can appear in $(*_{\mathcal{T}'})$. For any such variable x_i, since $\mathcal{T}'$ is a connected component of $\mathcal{G}$, it must contain both or neither of the two equations that x_i appears in. If $\mathcal{T}'$ contains both, then x_i does not appear in any term on the right-hand side of (16); if $\mathcal{T}'$ contains neither, then x_i does not appear in any term on the right-hand side of (15). So in either case, x_i does not appear in $(*'_\mathcal{T})$.

This means at most two variables can appear in $(*_{\mathcal{T}'})$ (namely, the ones that appear in more than two equations). But $*_{\mathcal{T}'}$ cannot be 0 (as the equations in $\mathcal{T}'$ are linearly independent), and its coefficients must sum to 0; this means it must be of the form $\varepsilon x_{i_1} - \varepsilon x_{i_2} = 0$ for some $\varepsilon \neq 0$ and $i_1 \neq i_2$, contradicting the assumption that $\mathcal{T}$ is valid.

Since we have obtained a contradiction, $\mathcal{G}$ must in fact be connected, which proves (ii).

Finally, (ii) implies that every variable is contained in an even number of $(*_1)$, ..., $(*_t)$, $(*_\mathcal{T})$. This eliminates the possibility that two variables appear in exactly 3 equations, finishing the proof of (i). □

Lemma 3.8. *Let $\mathcal{T}$ be a 2-good collection of linearly independent difference equalities. Then there is at most one difference equality that $\mathcal{T}$ minimally implies.*

Proof. If $\mathcal{T}$ minimally implies a difference equality $(*_\mathcal{T})$, then by Lemma 3.7, every variable appears an even number of times among $\mathcal{T}$ and $(*_\mathcal{T})$, so the variables appearing in $(*_\mathcal{T})$ are precisely the ones that appear in an odd number of equations in $\mathcal{T}$. Then if $\mathcal{T}$ minimally implied two distinct difference equalities, these difference equalities would have to involve the same 4 variables, contradicting Lemma 3.2. □

3.3. **Proof of Lemma 3.3.** In this subsection, we prove Lemma 3.3 (which analyzes the 2-good collections $\mathcal{S}$ that achieve equality or near-equality in Lemma 2.9). Recall that we proved Lemma 2.9 by drawing 'boxes' around all subsets of $\mathcal{S}$ that minimally imply some difference equality involving x_i, and showing that each connected 'blob' formed by these boxes must involve reasonably few variables. The idea of our proof of Lemma 3.3 is essentially that if $\mathcal{S}$ is 2-good, then Lemma 3.7 heavily restricts the structure of these boxes and blobs — it will immediately imply that all boxes have size 1 or 3, and we will later show that the boxes of size 3 must be disjoint. This additional structure makes it much simpler to analyze when equality or near-equality holds in the bounds used to prove Lemma 2.9.

First, by Lemma 3.7(i), if $\{(*_1), \ldots, (*_t)\}$ is a 2-good collection of linearly independent difference equalities involving x_i that minimally implies a difference equality $(*)$ also involving x_i, then t must be 1 or 3, since x_i appears in all $t + 1$ equations. (In the terminology used in the proof of Lemma 2.9, this states that all boxes have size 1 or 3.) The case where $t = 1$ is straightforward to analyze — $(*)$ must be the same equation as $(*_1)$. We now analyze the case where $t = 3$. We call such a set $\{(*_1), (*_2), (*_3)\}$ a 3-*implication at i* (or simply a 3-*implication*, if we do not need to specify i), and we say it *produces* the difference equality $(*)$. Note that by Lemma 3.8, each 3-implication produces only one difference equality.

Lemma 3.9. *Let $\mathcal{T}$ be a 3-implication and $(*)$ the difference equality it produces. Let $\mathcal{G}_3(\mathcal{T})$ be the graph with vertex set $\mathcal{T} \cup \{(*)\}$ where two equations are adjacent if they are difference-aligned. Then $\mathcal{G}_3(\mathcal{T})$ must be a 4-vertex path.*

Proof. Let $\mathcal{T} = \{(*_1), (*_2), (*_3)\}$, and suppose that $\mathcal{T}$ is a 3-implication at 1 involving the 7 variables x_1, ..., x_7 (by Lemma 3.7(i), any 3-implication involves exactly 7 variables). Let $*_1, *_2, *_3, *$ be the contents of $(*_1), (*_2), (*_3), (*)$; we may assume that x_1 has a coefficient of $+1$ in each. Then by Lemma 3.7(ii), we can write $* = \varepsilon_1 *_1 + \varepsilon_2 *_2 + \varepsilon_3 *_3$ for some $\varepsilon_1, \varepsilon_2, \varepsilon_3 \in \{\pm 1\}$. Exactly two of $\varepsilon_1, \varepsilon_2, \varepsilon_3$ must be $+1$, so we can assume without loss of generality that

$$ (17) \qquad \qquad * = *_1 + *_2 - *_3. $$

First note that each of the 6 variables x_2, ..., x_7 must be shared by exactly one pair of $(*_1), (*_2), (*_3)$, $(*)$; meanwhile, each of the $\binom{4}{2} = 6$ pairs of equations shares at most one such variable by Lemma 3.2 (since the pair already shares x_1). This means each pair of equations must share *exactly* one variable.

Then the variable shared by $(*_1)$ and $(*_2)$ must have opposite coefficients in $*_1$ and $*_2$, so we can assume without loss of generality that

$$ *_1 = x_1 - x_2 - x_3 + x_4 $$
$$ *_2 = x_1 + x_2 \qquad \quad - x_5 - x_6. $$

Then one of x_5 and x_6 must appear in $(*_3)$, and the other must appear in $(*)$. Both must appear with opposite sign as x_1 in their respective equations in order to cancel out of (17), so $(*_2)$ must be difference-aligned with both $(*_3)$ and $(*)$. For the same reason, $(*_1)$ is difference-aligned with whichever of $(*_3)$ and $(*)$ involves x_3 (and is not difference-aligned with either of the other two equations). Finally, $(*_3)$ and $(*)$ cannot be difference-aligned for the same reason that $(*_1)$ and $(*_2)$ cannot be — the shared variable x_7 must have opposite coefficients in $*_3$ and $*_4$. So $\mathcal{G}_3(\mathcal{T})$ is a 4-vertex path. □

Figure 3 shows an example of the graph $\mathcal{G}_3(\mathcal{T})$. In fact, by slightly extending the argument in our proof of Lemma 3.9, one can show that *any* 3-implication must look like the one in Figure 3, up to swapping which 3-equation subset of the graph we consider to be $\mathcal{T}$ (and, of course, renaming the variables).
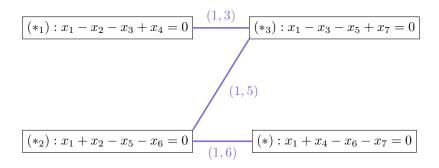
FIGURE 3. This figure shows an example of a 3-implication $\mathcal{T}$ at 1 and the corresponding graph $\mathcal{G}_3(\mathcal{T})$; here each edge of $\mathcal{G}_3(\mathcal{T})$ is labelled with the pair $(1, j)$ at which its endpoints are difference-aligned.

Lemma 3.9 describes how an individual 3-implication can look; the following lemma describes how two 3-implications can interact with each other.

Lemma 3.10. *Fix $i \in \{1, \ldots, k\}$. Let $\mathcal{S}$ be a 2-good collection of linearly independent difference equalities involving x_i, and suppose that $\mathcal{T}_1$ and $\mathcal{T}_2$ are distinct 3-implications contained in $\mathcal{S}$. Then $\mathcal{T}_1$ and $\mathcal{T}_2$ are disjoint (i.e., they have no equations in common).*

Proof. Assume that $\mathcal{T}_1$ and $\mathcal{T}_2$ are not disjoint, so they have 1 or 2 equations in common.

First suppose they have exactly 1 equation in common. Each of $\mathcal{T}_1$ and $\mathcal{T}_2$ involves 7 variables, and the shared equation involves 4; this means $\mathcal{T}_1 \cup \mathcal{T}_2$ involves at most $7 + 7 - 4 = 10$ variables. Then $\mathcal{T}_1 \cup \mathcal{T}_2$ contains 5 linearly independent equations on at most 10 variables; this contradicts the assumption that $\mathcal{S}$ is 2-light (by Remark 2.2).

Now suppose they have exactly 2 equations in common. Then by Lemma 3.2, the 2 shared equations together involve at least 6 variables; this means $\mathcal{T}_1 \cup \mathcal{T}_2$ involves at most $7 + 7 - 6 = 8$ variables. But then $\mathcal{T}_1 \cup \mathcal{T}_2$ contains 4 linearly independent equations on at most 8 variables; this again contradicts the assumption that $\mathcal{S}$ is 2-light. $\square$

We now have all the pieces needed to prove Lemma 3.3. We first prove the following slightly more general statement, from which it will be easy to deduce Lemma 3.3.

Lemma 3.11. *Fix $i \in \{1, \ldots, k\}$, and let $\mathcal{S}$ be a 2-good collection consisting of s linearly independent difference equalities that all involve x_i. Then we can find some $\mathcal{S}'$, also consisting of s linearly independent difference equalities that all involve x_i, such that $\mathcal{S}'$ is equivalent to $\mathcal{S}$ and such that for each index $j \neq i$, $\mathcal{S}$ i-certifies j if and only if x_i and x_j appear with opposite sign in an equation in $\mathcal{S}'$.*

Proof. Note that every difference equality involving x_i implied by $\mathcal{S}$ either is already present in $\mathcal{S}$ or is produced by a 3-implication. Furthermore, the 3-implications in $\mathcal{S}$ are all disjoint (by Lemma 3.10), and each produces only one difference equality (by Lemma 3.8).

We now define $\mathcal{S}'$ in the following way:

- For each equation $(*)$ in $\mathcal{S}$ *not* contained in any 3-implication, we simply place $(*)$ into $\mathcal{S}'$.
- For each 3-implication $\mathcal{T} \subseteq \mathcal{S}$, let $(*_\mathcal{T})$ be the equation it produces. By Lemma 3.9, the graph $\mathcal{G}_3(\mathcal{T})$ on $\mathcal{T} \cup \{(*_\mathcal{T})\}$ is a 4-vertex path by Lemma 3.9; we consider one vertex of degree 2, and place the three *other* equations in $\mathcal{T} \cup \{(*_\mathcal{T})\}$ into $\mathcal{S}'$.

An example of this process is shown in Figure 4.

Then it is clear that $\mathcal{S}'$ consists of s linearly independent difference equalities involving x_i, and that $\mathcal{S}'$ is equivalent to $\mathcal{S}$ (note that if $\mathcal{T}$ is a 3-implication producing $(*_\mathcal{T})$, then all 3-equation subsets of $\mathcal{T} \cup \{(*_\mathcal{T})\}$ are equivalent).

Finally, we show that $\mathcal{S}$ i-certifies an index j if and only if x_i and x_j appear with opposite sign in some equation in $\mathcal{S}'$. By definition, $\mathcal{S}$ i-certifies j if and only if x_i and x_j appear with opposite sign in some difference equality implied by $\mathcal{S}$. Of course every equation in $\mathcal{S}'$ is implied by $\mathcal{S}$, so if x_i and x_j appear with opposite sign in an equation in $\mathcal{S}'$, then $\mathcal{S}$ i-certifies j. Conversely, if $(*)$ is a difference equality implied by

$$\mathcal{S} \qquad\qquad\qquad\qquad \mathcal{S}'$$

$$(x_1 + x_4 - x_6 - x_7 = 0)$$

$x_1 - x_2 - x_3 + x_4 = 0$		$x_1 - x_2 - x_3 + x_4 = 0$
$x_1 + x_2 - x_5 - x_6 = 0$	$\longrightarrow$	$x_1 - x_3 - x_5 + x_7 = 0$
$x_1 - x_3 - x_5 + x_7 = 0$		$x_1 + x_4 - x_6 - x_7 = 0$

$$x_1 + x_6 - x_8 - x_9 = 0 \qquad\longrightarrow\qquad x_1 + x_6 - x_8 - x_9 = 0$$

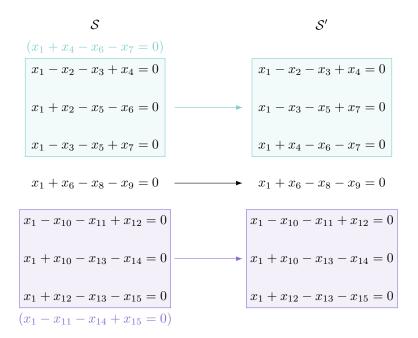| $x_1 - x_{10} - x_{11} + x_{12} = 0$		$x_1 - x_{10} - x_{11} + x_{12} = 0$
$x_1 + x_{10} - x_{13} - x_{14} = 0$	$\longrightarrow$	$x_1 + x_{10} - x_{13} - x_{14} = 0$
$x_1 + x_{12} - x_{13} - x_{15} = 0$		$x_1 + x_{12} - x_{13} - x_{15} = 0$

$$(x_1 - x_{11} - x_{14} + x_{15} = 0)$$

FIGURE 4. This figure shows an example of a possible collection $\mathcal{S}$ (on the left-hand side) and one way to choose the corresponding $\mathcal{S}'$ (on the right-hand side), with $i = 1$. The 3-implications in $\mathcal{S}$ are shown in boxes and labelled with the difference equalities they produce (shown in parentheses — the equations in parentheses are not part of $\mathcal{S}$). In this example, for the first 3-implication (boxed in blue) we leave out $x_1 + x_2 - x_5 - x_6 = 0$, and for the second 3-implication (boxed in purple) we leave out $x_1 - x_{11} - x_{14} + x_{15} = 0$.

$\mathcal{S}$ that is *not* in $\mathcal{S}'$, then it must be a vertex of degree 2 in $\mathcal{G}_3(\mathcal{T})$ for some 3-implication $\mathcal{T} \subseteq \mathcal{S}$. This means it is difference-aligned with two other equations in $\mathcal{G}_3(\mathcal{T})$, so both variables x_j that appear with opposite sign as x_i in $(*)$ must also appear with opposite sign as x_i in one of the other three equations in $\mathcal{G}_3(\mathcal{T})$. (Note that the two equations difference-aligned with $(*)$ in $\mathcal{G}_3(\mathcal{T})$ cannot both be difference-aligned with $(*)$ at the *same* pair (i, j), as then they would be difference-aligned with each other and $\mathcal{G}_3(\mathcal{T})$ would contain a triangle.) All three of these equations are present in $\mathcal{S}'$, so $\mathcal{S}'$ has the desired property. $\qquad\square$

Proof of Lemma 3.3. Let $\mathcal{S}'$ be the result of applying Lemma 3.11 to $\mathcal{S}$. Then it suffices to verify (i) and (ii), given that $\mathcal{S}$ i-certifies j if and only if x_i and x_j appear with opposite sign in some equation in $\mathcal{S}'$.

For (i), we are given that there are exactly $2s$ indices j such that x_i and x_j appear with opposite sign in some equation in $\mathcal{S}'$. Since each of the s equations in $\mathcal{S}'$ has exactly two variables that appear with opposite sign as x_i, this means for each such j, the variable x_j appears with opposite sign in only one equation in $\mathcal{S}'$, and therefore no two equations in $\mathcal{S}'$ are difference-aligned. (Recall that by Lemma 3.2, since any two equations in $\mathcal{S}'$ share x_i, if they are difference-aligned then they must be difference-aligned at (i, j) for some j, meaning that some variable x_j must appear with opposite sign as x_i in both.)

Similarly, for (ii), we are given that there are exactly $2s - 1$ indices j such that x_i and x_j appear with opposite sign in some equation in $\mathcal{S}'$. This means one such variable x_j must appear with opposite sign as x_i in exactly two equations in $\mathcal{S}'$, and the rest must appear with opposite sign as x_i in only one; so exactly one pair of equations in $\mathcal{S}'$ is difference-aligned. $\qquad\square$

3.4. **Proof of Lemma 3.4.** In this subsection, we prove Lemma 3.4. This proof will involve analyzing collections of linearly independent difference equalities involving x_i that minimally imply some difference equality *not* involving x_i. Similarly to in the previous subsection, the main idea is that Lemma 3.7 heavily restricts how such a collection can look, which will allow us to bound the number of such collections.

First, by Lemma 3.7(i), if $\{(*_1), \ldots, (*_t)\}$ is a 2-good collection of linearly independent difference equalities involving x_i that minimally implies some difference equality $(*)$ *not* involving x_i, then t must be 2 or 4 (as x_i appears in exactly t of these equations). We call such a collection $\{(*_1), \ldots, (*_t)\}$ a 2-*implication at* i

(or simply a 2-*implication*, if we do not need to specify i) if $t = 2$, and a 4-*implication at i* (or simply a 4-*implication*) if $t = 4$; and we say it *produces* the difference equality (∗). Again by Lemma 3.8, any 2-implication or 4-implication produces only one difference equality.

First, 2-implications are easy to understand.

Lemma 3.12. *The 2 equations of any 2-implication are either difference-aligned or sum-aligned.*

Proof. Let $\{(∗_1), (∗_2)\}$ be a 2-implication at i, and let (∗) be the difference equality it produces. Let $∗_1$, $∗_2$, ∗ be the contents of $(∗_1)$, $(∗_2)$, (∗); we can assume without loss of generality that x_i has a coefficient of $+1$ in $∗_1$ and $∗_2$, and that

$$∗ = ∗_1 - ∗_2.$$

Then $∗_1$ and $∗_2$ must share some variable other than x_i (otherwise (∗) would involve 7 variables), and this variable must have the same coefficient in both; this means $(∗_1)$ and $(∗_2)$ must be either difference-aligned (if this coefficient is -1) or sum-aligned (if this coefficient is $+1$). □

We now turn to 4-implications. To prove Lemma 3.4, we only need to consider 4-implications in which at most one pair of equations is difference-aligned (since we assume that the collection $\mathcal{S}$ in Lemma 3.4 contains at most one pair of difference-aligned equations).

Lemma 3.13. *Let $\mathcal{T}$ be a 4-implication such that at most one pair of equations in $\mathcal{T}$ is difference-aligned. Then exactly one pair of equations in $\mathcal{T}$ is difference-aligned, and exactly one pair is sum-aligned.*

Proof. Let $\mathcal{T} = \{(∗_1), (∗_2), (∗_3), (∗_4)\}$, and let (∗) be the difference equality produced by $\mathcal{T}$. Suppose that $\mathcal{T}$ is a 4-implication at 1 involving the 9 variables x_1, …, x_9 (by Lemma 3.7(i), any 4-implication involves exactly 9 variables). Let $∗_1$, $∗_2$, $∗_3$, $∗_4$, ∗ be the contents of $(∗_1)$, $(∗_2)$, $(∗_3)$, $(∗_4)$, (∗); we can assume without loss of generality that the coefficient of x_1 in each of $∗_1$, $∗_2$, $∗_3$, $∗_4$ is $+1$. Then by Lemma 3.7(ii) we have $∗ = \varepsilon_1 ∗_1 + \cdots + \varepsilon_4 ∗_4$ for some $\varepsilon_1, \ldots, \varepsilon_4 \in \{\pm 1\}$; exactly two of these coefficients must be $+1$, so we can assume without loss of generality that

$$∗ = ∗_1 + ∗_2 - ∗_3 - ∗_4.$$

First we claim that $∗_1$ and $∗_2$ must share exactly one variable other than x_1. To see this, by Lemma 3.2 they cannot share more than one such variable; now assume for contradiction that they share none. Then we can assume without loss of generality that

$$∗_1 = x_1 - x_2 - x_3 + x_4$$
$$∗_2 = x_1 \qquad\qquad - x_5 - x_6 + x_7.$$

Then at least two of x_2, x_3, x_5, and x_6 must appear in $∗_3$ or $∗_4$ with coefficient -1, as any of these variables that does not do so will appear in ∗ with coefficient -1 (and ∗ can only have two variables with coefficient -1). But this means at least two pairs of $(∗_1)$, $(∗_2)$, $(∗_3)$, $(∗_4)$ are difference-aligned, which is a contradiction.

So $∗_1$ and $∗_2$ share exactly one variable other than x_1, and this variable must have coefficient -1 in one and $+1$ in the other; so we can assume without loss of generality that

$$∗_1 = x_1 - x_2 - x_3 + x_4$$
$$∗_2 = x_1 + x_2 \qquad\qquad - x_5 - x_6.$$

As before, at most two of x_3, x_5, and x_6 can appear in ∗ (as each would have coefficient -1), while any of these variables that does *not* appear in ∗ must appear in $∗_3$ or $∗_4$ with coefficient -1, producing a difference-aligned pair of equations. Since we have at most one difference-aligned pair, this means *exactly* two of x_3, x_5, and x_6 must appear in ∗, and there is exactly one difference-aligned pair of equations (the pair sharing the one variable among x_3, x_5, and x_6 that does not appear in ∗).

Finally, $∗_3$ and $∗_4$ must also share exactly one variable other than x_1 (for the same reason that $∗_1$ and $∗_2$ do); let this variable be x_7, and assume that x_7 has coefficient -1 in $∗_3$ and $+1$ in $∗_4$. Then x_8 and x_9 must both appear in ∗ (if one did not, then it would also be shared by $∗_3$ and $∗_4$). Since only 4 variables appear in ∗, this means x_4 does not; so it must appear with coefficient $+1$ in $∗_3$ (it cannot appear with coefficient $+1$ in $∗_4$, as x_7 does). So there is exactly one sum-aligned pair, namely $(∗_1)$ and $(∗_3)$. □

Example 3.14. We illustrate Lemma 3.13 for two examples of 4-implications at 1; as in the proof of Lemma 3.13, we let our 4-implication be $\{(∗_1), (∗_2), (∗_3), (∗_4)\}$ and the difference equality it produces be (∗), and we let the contents of $(∗_1)$, $(∗_2)$, $(∗_3)$, $(∗_4)$, (∗) be $∗_1$, $∗_2$, $∗_3$, $∗_4$, ∗.

One example of a 4-implication with only one difference-aligned pair of equations is given by

$$
\begin{array}{llllll}
*_1 = x_1 - x_2 - x_3 + x_4 \\
*_2 = x_1 + x_2 & & & - x_5 - x_6 \\
*_3 = x_1 & & + x_4 - x_5 & & - x_7 \\
*_4 = x_1 & & & & + x_7 - x_8 - x_9 \\
\hline
\\
* = & & - x_3 & & - x_6 & + x_8 + x_9
\end{array}
$$

(where we have $* = *_1 + *_2 - *_3 - *_4$). Here the one difference-aligned pair is $\{(*_2), (*_3)\}$, and the one sum-aligned pair is $\{(*_1), (*_3)\}$.

Another example of a 4-implication with only one difference-aligned pair of equations is

$$
\begin{array}{llllll}
*_1 = x_1 - x_2 - x_3 + x_4 \\
*_2 = x_1 + x_2 & & & - x_5 - x_6 \\
*_3 = x_1 & & + x_4 & & - x_7 - x_8 \\
*_4 = x_1 & & - x_5 & & + x_7 & - x_9 \\
\hline
\\
* = & & - x_3 & & - x_6 & + x_8 + x_9
\end{array}
$$

(where we again have $* = *_1 + *_2 - *_3 - *_4$). Here the one difference-aligned pair is $\{(*_2), (*_4)\}$, and the one sum-aligned pair is $\{(*_1), (*_3)\}$.

In fact, by slightly extending the arguments in Lemmas 3.13 and 3.16, it can be shown that any 4-implication with only one difference-aligned pair of equations must be of one of these two forms.

Lemma 3.13 describes how an individual 4-implication can look; the next three lemmas analyze how 4-implications can interact with each other.

Lemma 3.15. *Fix $i \in \{1, \ldots, k\}$, and let $\mathcal{S}$ be a 2-good collection of linearly independent difference equalities involving x_i such that only one pair of equations in $\mathcal{S}$ is difference-aligned. Suppose that $\mathcal{T}_1$ and $\mathcal{T}_2$ are distinct 4-implications contained in $\mathcal{S}$. Then $\mathcal{T}_1$ and $\mathcal{T}_2$ share exactly 3 equations, which involve exactly 7 variables.*

Proof. First, by Lemma 3.13, $\mathcal{T}_1$ and $\mathcal{T}_2$ must both contain the one difference-aligned pair of equations in $\mathcal{S}$; this means they have either 2 or 3 equations in common.

First assume for contradiction that they have 2 equations in common. Then by Lemma 3.2, these 2 equations together involve at least 6 variables; since $\mathcal{T}_1$ and $\mathcal{T}_2$ each involves 9 variables, this means $\mathcal{T}_1 \cup \mathcal{T}_2$ involves at most $9 + 9 - 6 = 12$ variables. But then $\mathcal{T}_1 \cup \mathcal{T}_2$ consists of 6 linearly independent equations on at most 12 variables, contradicting the assumption that $\mathcal{S}$ is 2-light (by Remark 2.2).

So $\mathcal{T}_1$ and $\mathcal{T}_2$ have 3 equations in common, and these 3 equations must involve at least 7 variables (again by the assumption that $\mathcal{S}$ is 2-light and Remark 2.2). If they involved at least 8 variables, then $\mathcal{T}_1 \cup \mathcal{T}_2$ would consist of 5 linearly independent equations on at most $9 + 9 - 8 = 10$ variables, again contradicting the assumption that $\mathcal{S}$ is 2-light; so they must involve *exactly* 7. $\square$

Lemma 3.16. *Fix $i \in \{1, \ldots, k\}$, and let $\mathcal{T}$ be a 4-implication at i such that only one pair of equations in $\mathcal{T}$ is difference-aligned. Suppose that there is some 3-equation subset $\mathcal{T}' \subseteq \mathcal{T}$ that involves exactly 7 variables. Then there is only one such subset, and this subset $\mathcal{T}'$ forms a 3-implication at i producing some difference equality (involving x_i) that is sum-aligned with the fourth equation.*

For example, in the first 4-implication in Example 3.14, $\{(*_1), (*_2), (*_3)\}$ is such a subset; the second 4-implication in Example 3.14 has no such subset.

Proof. Let $\mathcal{T} = \{(*_1), (*_2), (*_3), (*_4)\}$, let $(*_\mathcal{T})$ be the difference equality it produces, and let $*_1, *_2, *_3, *_4, *_\mathcal{T}$ be the contents of $(*_1), (*_2), (*_3), (*_4), (*_\mathcal{T})$. By Lemma 3.7(ii), we can write

$$
*_\mathcal{T} = \varepsilon_1 *_1 + \cdots + \varepsilon_4 *_4
$$

for some $\varepsilon_1, \ldots, \varepsilon_4 \in \{\pm 1\}$.

Assume that $\mathcal{T}' = \{(*_1), (*_2), (*_3)\}$ involves 7 variables. Let

$$*_{\mathcal{T}'} = \varepsilon_1 *_1 + \varepsilon_2 *_2 + \varepsilon_3 *_3 = *_{\mathcal{T}} - \varepsilon_4 *_4,$$

and let $(*_{\mathcal{T}'})$ be the equation $*_{\mathcal{T}'} = 0$.

We claim that $(*_{\mathcal{T}'})$ is a difference equality involving x_i. To see this, first note that $*_1$, $*_2$, and $*_3$ together contain $3 \cdot 3 = 9$ 'slots' for variables other than x_i, so since they contain 6 variables other than x_i and each appears once or twice, 3 of these variables must appear once and 3 must appear twice. The 2 variables that do not appear in $*_1$, $*_2$, $*_3$ clearly cannot appear in $*_{\mathcal{T}'}$; meanwhile, the 3 variables that appear twice in $*_1$, $*_2$, $*_3$ cannot appear in $*_{\mathcal{T}}$ or $*_4$, and therefore cannot appear in $*_{\mathcal{T}'}$ either. On the other hand, the 3 variables that appear once in $*_1$, $*_2$, $*_3$, as well as x_i, must each appear in exactly one of $*_{\mathcal{T}}$ and $*_4$; so these four variables all appear in $*_{\mathcal{T}'}$ with coefficient ± 1. Finally, the coefficients of $*_{\mathcal{T}'}$ must sum to 0; this means $(*_{\mathcal{T}'})$ is a difference equality.

So $\{(*_1), (*_2), (*_3)\}$ is a 3-implication at i producing $(*_{\mathcal{T}'})$. By Lemma 3.10, this means no other subset can contain 7 variables (as then it would also form a 3-implication, so we would have two 3-implications that are not disjoint).

Finally, we need to show that $(*_{\mathcal{T}'})$ is sum-aligned with $(*_4)$. First note that $*_{\mathcal{T}} = *_{\mathcal{T}'} + \varepsilon_4 *_4$, so $\{(*_{\mathcal{T}'}), (*_4)\}$ is a 2-implication; by Lemma 3.12, this means $(*_{\mathcal{T}'})$ and $(*_4)$ are either difference-aligned or sum-aligned. But by Lemma 3.9 applied to $\mathcal{T}'$, the graph $\mathcal{G}_3(\mathcal{T}')$ is a 4-vertex path on $(*_1)$, $(*_2)$, $(*_3)$, $(*'_{\mathcal{T}})$. Since at most one pair of $(*_1)$, $(*_2)$, $(*_3)$ is difference-aligned by assumption (and therefore there is at most one edge between them), this means *exactly* one pair is, and $(*_{\mathcal{T}'})$ must be one of the vertices of degree 2 in this path — so both variables that appear with opposite sign as x_i in $(*_{\mathcal{T}'})$ must also appear with opposite sign as x_i in one of $(*_1)$, $(*_2)$, $(*_3)$. Then if $(*_4)$ were difference-aligned with $(*_{\mathcal{T}'})$, it would also be difference-aligned with one of $(*_1)$, $(*_2)$, $(*_3)$, producing a second difference-aligned pair among $(*_1)$, $(*_2)$, $(*_3)$, $(*_4)$. This is impossible, so $(*_{\mathcal{T}'})$ and $(*_4)$ must instead be sum-aligned. $\square$

Lemma 3.17. *Fix $i \in \{1, \ldots, k\}$, and let $\mathcal{S}$ be a 2-good collection of linearly independent difference equalities involving x_i such that at most one pair is difference-aligned. Suppose that $\mathcal{T}_1$, $\ldots$, $\mathcal{T}_r$ are distinct 4-implications contained in $\mathcal{S}$. Then there exist three equations present in all of them, and their fourth equations are all sum-aligned.*

Proof. If $r = 1$, the statement is vacuously true; now assume $r \geq 2$. Then by Lemma 3.15, each of $\mathcal{T}_2$, $\ldots$, $\mathcal{T}_r$ shares exactly 3 equations with $\mathcal{T}_1$, and these 3 equations involve exactly 7 variables. Furthermore, by Lemma 3.16, $\mathcal{T}_1$ only has one 3-equation subset $\mathcal{T}'$ involving exactly 7 variables, so it must share the same subset $\mathcal{T}'$ with each of $\mathcal{T}_2$, $\ldots$, $\mathcal{T}_r$.

Then each of $\mathcal{T}_1$, $\ldots$, $\mathcal{T}_r$ contains the 3-equation subset $\mathcal{T}'$. Furthermore, by Lemma 3.16, $\mathcal{T}'$ must form a 3-implication at i producing a difference equality $(*'_{\mathcal{T}})$ involving x_i, and the fourth equations of $\mathcal{T}_1$, $\ldots$, $\mathcal{T}_r$ must all be sum-aligned with $(*_{\mathcal{T}'})$. This means these fourth equations are also sum-aligned with each other (recall that since these equations and $(*_{\mathcal{T}'})$ all share x_i, by Lemma 3.2 if two are sum-aligned then they must be sum-aligned at (i, j) for some j). $\square$

We are now ready to prove Lemma 3.4.

Proof of Lemma 3.4. Write down all the equations in $\mathcal{S}$, and split them into *stacks* based on which variable occurs with the same sign as x_i (in other words, two equations are sum-aligned if and only if they are in the same stack). Suppose that we have r stacks with sizes s_1, $\ldots$, s_r, so that $s_1 + \cdots + s_r = s$.

Every difference equality implied by $\mathcal{S}$ that does not involve x_i must be produced by a 2-implication or a 4-implication at i. Furthermore, each 2-implication and each 4-implication can only produce one difference equality. So in order to bound the number of difference equalities implied by $\mathcal{S}$ that do not involve x_i, it suffices to bound the total number of 2-implications and 4-implications at i contained in $\mathcal{S}$.

First, we prove (iii): the given condition means that there is only one stack, and therefore there are exactly $\binom{s}{2}$ 2-implications formed by sum-aligned pairs of equations. Meanwhile, since two equations cannot be both sum-aligned and difference-aligned, there are no difference-aligned pairs of equations in $\mathcal{S}$; by Lemma 3.13, this also means there are no 4-implications in $\mathcal{S}$. So the total number of 2-implications and 4-implications is precisely $\binom{s}{2}$, as desired.

Next, we prove (i): now we are given that there are at least two stacks, and that no two equations in $\mathcal{S}$ are difference-aligned. Then for the same reason as in the previous case, the only 2-implications and

4-implications contained in $\mathcal{S}$ are the 2-implications formed by sum-aligned pairs, of which there are exactly

$$\binom{s_1}{2} + \cdots + \binom{s_r}{2}.$$

Since the function $x \mapsto \binom{x}{2}$ is convex and there are at least two stacks (of sizes at least 1), this is at most

$$\binom{s-1}{2} + \binom{1}{2} = \binom{s-1}{2}.$$

So the total number of 2-implications and 4-implications is at most $\binom{s-1}{2}$, as desired.

Finally, we prove (ii): we are given that exactly one pair of equations in $\mathcal{S}$ is difference-aligned; since two equations cannot be both difference-aligned and sum-aligned, this means there are at least two stacks.

First consider the case where $\mathcal{S}$ does not contain any 4-implications. Then $\mathcal{S}$ contains $\binom{s_1}{2} + \cdots + \binom{s_r}{2}$ 2-implications in which the two equations are sum-aligned, and one where the two equations are difference-aligned. So the number of 2-implications contained in $\mathcal{S}$ is

$$\binom{s_1}{2} + \cdots + \binom{s_r}{2} + 1 \le \binom{s-1}{2} + 1.$$

Now consider the case where $\mathcal{S}$ *does* contain a 4-implication. By Lemma 3.13, any 4-implication must span exactly 3 stacks (as it contains exactly one sum-aligned pair), so there are at least 3 nonempty stacks. Furthermore, by Lemma 3.17, there exist three equations present in all 4-implications, and there is some fixed stack — without loss of generality, the stack of size s_1 — that the fouth equation of each 4-implication belongs to. Then the number of 4-implications is at most s_1. So the total number of 2-implications and 4-implications is at most

$$\binom{s_1}{2} + \cdots + \binom{s_r}{2} + 1 + s_1 = \binom{s_1+1}{2} + \binom{s_2}{2} + \cdots + \binom{s_r}{2} + 1.$$

Since $(s_1 + 1) + s_2 + \cdots + s_r = s + 1$ and there are at least three stacks (of sizes at least 1), this is at most

$$\binom{s-1}{2} + \binom{1}{2} + \binom{1}{2} + 1 = \binom{s-1}{2} + 1.$$

So in either case, the total number of 2-implications and 4-implications is at most $\binom{s-1}{2}+1$, as desired. $\square$

Thus we have shown the technical lemmas we used to prove Lemma 3.1, which completes the proof of Proposition 1.7.

4. Proof of Theorem 1.8 (family of lower bounds on $g(n,k,\ell)$)

In this section, we prove Theorem 1.8, our lower bounds on $g(n, k, \ell)$ for certain values of ℓ. The idea of the proof is as follows: we show that if $|A - A|$ is too small, then we can find a certain k-element structure A' in A that satisfies $|A' - A'| \le \ell - 1$ for the value of ℓ in Theorem 1.8, which contradicts the (k, ℓ)-local property. Finding this k-element structure will involve finding a sum or difference that is repeated many times. For this, we will need the following lemma. Clearly, a set with small difference set must have a difference that is repeated many times; the following lemma states that such a set must also have a *sum* that is repeated many times.

Lemma 4.1. *If $A \subseteq \mathbb{R}$ has n elements, then*

$$\max_{c \in \mathbb{R}} \left(\#\{(a, b) \in A^2 \mid a + b = c\} \right) \cdot \#\{a - b \mid (a, b) \in A^2\} \ge n^2.$$

Note that $\{a - b \mid (a, b) \in A^2\}$ includes *all* differences between two elements of A, not just the positive differences; in particular, we have $\#\{a - b \mid (a, b) \in A^2\} = 2|A - A| + 1$.

Proof. Let $m = \max_{c \in \mathbb{R}} \left(\#\{(a, b) \in A^2 \mid a + b = c\} \right)$ and $d = \#\{a - b \mid (a, b) \in A^2\}$. Let

$$S = \{(a_1, a_2, a_3, a_4) \in A^4 \mid a_1 - a_2 = a_3 - a_4\}.$$

We will count $|S|$ (the additive energy of A) in two ways, one in terms of m and the other in terms of d.

First note that any $(a_1, a_2, a_3, a_4) \in S$ satisfies $a_1 + a_4 = a_2 + a_3$. So if we try to choose an element of S by first choosing a_1 and a_4 (which can be done in n^2 ways), then there are at most m ways to choose a_2 and a_3. This means $|S| \le n^2 m$.

Now, suppose that the d differences $a - b$ over all n^2 pairs $(a, b) \in A^2$ occur with multiplicities x_1, ..., x_d respectively, so that $x_1 + \cdots + x_d = n^2$. Then by first fixing some difference and then choosing (a_1, a_2) and (a_3, a_4) to be two pairs achieving that difference, we have $|S| = x_1^2 + \cdots + x_d^2$. By the Cauchy–Schwarz inequality, this means $|S| \geq \frac{n^4}{d}$.

Combining these two bounds gives $\frac{n^4}{d} \leq |S| \leq n^2 m$, so $md \geq n^2$.   □

The next lemma states that we can find a certain structure in any set A with sufficiently small $|A - A|$. Intuitively, this structure consists of $2s$ congruent affine $(t - 1)$-cubes, whose centers form s pairs that all have the same sum.

Lemma 4.2. *Let s and t be positive integers. Suppose that $n > 1$ and that A is an n-element set such that*

$$|A - A| \leq \left(\frac{n}{8s}\right)^{1 + \frac{1}{2^t - 1}}.$$

Then there exist $d_1, \ldots, d_{t-1} \in \mathbb{R}_{>0}$ and $a_1, \ldots, a_{2s} \in A$ such that $a_1 + a_{s+1} = a_2 + a_{s+2} = \cdots = a_s + a_{2s}$, and the $2^t \cdot s$ numbers

$$a_i + \varepsilon_1 d_1 + \cdots + \varepsilon_{t-1} d_{t-1}$$

over all $1 \leq i \leq 2s$ and $\varepsilon_1, \ldots, \varepsilon_{t-1} \in \{0, 1\}$ are all distinct and belong to A.

Proof. We use induction on t. First note that since $n > 1$, we must have $|A - A| \geq 1$.

For the base case $t = 1$ (in which case we are given $|A - A| \leq \frac{n^2}{64s^2}$ and we only need to find distinct $a_1, \ldots, a_{2s} \in A$ such that $a_1 + a_{s+1} = \cdots = a_s + a_{2s}$), we have

$$\#\{a - b \mid (a, b) \in A^2\} = 2|A - A| + 1 \leq 3|A - A| \leq \frac{3n^2}{64s^2} \leq \frac{n^2}{3s},$$

so by Lemma 4.1 there must exist some $c \in \mathbb{R}$ for which $\#\{(a, b) \in A^2 \mid a + b = c\} \geq 3s$. Fix some $c \in \mathbb{R}$ with this property. Then at most one of these pairs (a, b) can have $a = b$, and of the remaining pairs, any two are either disjoint or the same up to order (i.e., $(a_2, b_2) = (b_1, a_1)$). So there must exist at least $\frac{3s-1}{2} \geq s$ disjoint pairs of distinct elements $(a, b) \in A^2$ with $a + b = c$, and letting (a_1, a_{s+1}), ..., (a_s, a_{2s}) be s of these pairs gives the desired result.

For the inductive step, first note that the given condition on $|A - A|$ can be rewritten as

$$(18) \qquad\qquad |A| \geq 8s \cdot |A - A|^{1 - \frac{1}{2^t}}.$$

This form of the condition will be more useful when we apply the inductive hypothesis.

Assume that $t \geq 2$ and that we have proven the statement for s and $t - 1$ (for all $n > 1$ and all n-element sets A with the stated property); we will then prove the statement for s and t. First, since there are $\binom{n}{2} \geq \frac{n^2}{4}$ total positive differences between pairs of elements in A (counting multiplicity), some difference must be repeated at least

$$\frac{\binom{n}{2}}{|A - A|} \geq \frac{n^2}{4 \cdot |A - A|} \geq \frac{64s^2 \cdot |A - A|^{2 - \frac{1}{2^t - 1}}}{4 \cdot |A - A|} = 16s^2 \cdot |A - A|^{1 - \frac{1}{2^t - 1}} \geq 16s \cdot |A - A|^{1 - \frac{1}{2^t - 1}}$$

times; let $d_{t-1} > 0$ be such a difference. For notational convenience, define

$$m = 8s \cdot |A - A|^{1 - \frac{1}{2^t - 1}},$$

so that there are at least $2m$ pairs $(a, b) \in A^2$ with $a - b = d_{t-1}$ (and $a \neq b$ in all such pairs). We claim that there are at least m such *disjoint* pairs. To see this, consider the graph on all such pairs (a, b) in which two pairs are adjacent if they have an element in common. This graph must be a collection of vertex-disjoint paths (as all vertices have degree at most 2, and there are no cycles), so we can obtain an independent set consisting of at least half the vertices (by taking alternating vertices on each path, starting with one endpoint), which provides at least m disjoint pairs $(a, b) \in A^2$ with $a - b = d_{t-1}$.

Now let A' be the set of values of a across these pairs, so that A' and $A' + d_{t-1}$ are disjoint and both contained in A, and

$$|A'| \geq m = 8s \cdot |A - A|^{1 - \frac{1}{2^t - 1}} \geq 8s \cdot |A' - A'|^{1 - \frac{1}{2^t - 1}}.$$

Then we can apply the inductive hypothesis to A', as A' satisfies the condition (18) for $t - 1$ (and we have $|A'| \geq m \geq 8s > 1$). This gives that we can find $a_1, \ldots, a_{2s} \in A'$ and $d_1, \ldots, d_{t-2} \in \mathbb{R}_{>0}$ such that $a_1 + a_{s+1} = \cdots = a_s + a_{2s}$ and the $2^{t-1} \cdot s$ numbers

$$a_i + \varepsilon_1 d_1 + \cdots + \varepsilon_{t-2} d_{t-2}$$

are all distinct and belong to A'. Then the $2^{t-1} \cdot s$ numbers

$$a_i + \varepsilon_1 d_1 + \cdots + \varepsilon_{t-2} d_{t-2} + d_{t-1}$$

are all distinct and belong to $A' + d_{t-1}$. Since A' and $A' + d_{t-1}$ are disjoint and both contained in A, all $2^t \cdot s$ numbers are distinct and belong to A, as desired. □

The following lemma provides an upper bound on the number of distinct differences present in a structure of the form given by Lemma 4.2.

Lemma 4.3. *Let $a_1, \ldots, a_{2s}, d_1, \ldots, d_{t-1}$ be fixed real numbers such that $a_1 + a_{s+1} = \cdots = a_s + a_{2s}$, and let $A' = \{a_i + \varepsilon_1 d_1 + \cdots + \varepsilon_{t-1} d_{t-1} \mid 1 \leq i \leq 2s \text{ and } \varepsilon_1, \ldots, \varepsilon_{t-1} \in \{0,1\}\}$. Then*

$$|A' - A'| \leq 3^{t-1} s^2 + \frac{3^{t-1} - 1}{2}.$$

Proof. Any difference $a - b$ for $(a, b) \in A^2$, including negative differences and zero, is of the form

$$(a_{i_1} - a_{i_2}) + \lambda_1 d_1 + \cdots + \lambda_{t-1} d_{t-1},$$

where $1 \leq i_1, i_2 \leq 2s$ and $\lambda_1, \ldots, \lambda_{t-1} \in \{-1, 0, 1\}$. The argument in Remark 2.10 (with $k = 2s$ and $d = s + 1$) shows that there are at most s^2 positive differences among $a_1, \ldots, a_{2s}$, and therefore at most $2s^2 + 1$ possible values of $a_{i_1} - a_{i_2}$ (including negative values and zero). Meanwhile, there are 3^{t-1} choices of values for $\lambda_1, \ldots, \lambda_{t-1}$; this means there are at most $3^{t-1}(2s^2 + 1)$ such differences. Since this includes negative differences and zero, this means

$$|A' - A'| \leq \frac{3^{t-1}(2s^2 + 1) - 1}{2} = 3^{t-1} s^2 + \frac{3^{t-1} - 1}{2}. \qquad \square$$

Together, Lemmas 4.2 and 4.3 immediately imply Theorem 1.8.

Proof of Theorem 1.8. Let $s = \frac{k}{2^t}$ and $\ell = 3^{t-1} s^2 + \frac{3^{t-1} + 1}{2}$; then we claim that for all n, we have

$$g(n, k, \ell) > \left(\frac{n}{8s}\right)^{1 + \frac{1}{2^t - 1}}.$$

Assume not; then there exists an n-element set A with $|A - A| \leq \left(\frac{n}{8s}\right)^{1 + \frac{1}{2^t - 1}}$ satisfying the (k, ℓ)-local property. But by Lemma 4.3 we can find a k-element subset $A' \subseteq A$ of the form

$$A' = \{a_i + \varepsilon_1 d_1 + \cdots + \varepsilon_{t-1} d_{t-1} \mid 1 \leq i \leq 2s \text{ and } \varepsilon_1, \ldots, \varepsilon_{t-1} \in \{0,1\}\}$$

for some $a_1, \ldots, a_{2s}, d_1, \ldots, d_{t-1}$ satisfying $a_1 + a_{s+1} = \cdots = a_s + a_{2s}$, and by Lemma 4.3 we must have $|A' - A'| \leq \ell - 1$, contradicting the assumption that A satisfies the (k, ℓ)-local property. □

5. DISCUSSION OF CONSEQUENCES

In this section, we deduce and discuss the more intuitive consequences on the behavior of $g(n, k, \ell)$ stated in the beginning of Subsection 1.4 (i.e., Corollaries 1.1, 1.2, 1.3, 1.4, and 1.5).

5.1. The quadratic threshold.
We begin by discussing our results on the quadratic threshold (i.e., Corollaries 1.1 and 1.2).

When k is even, Theorems 1.6 and 1.8 together immediately imply Corollary 1.1:

- Applying Theorem 1.6 with $c = 2$ gives that $g(n, k, \frac{k^2}{4}) = o(n^2)$.
- Applying Theorem 1.8 with $t = 1$ gives that $g(n, k, \frac{k^2}{4} + 1) = \Omega(n^2)$.

In particular, these two bounds together mean that the quadratic threshold is exactly $\frac{k^2}{4} + 1$.

Remark 5.1. We now comment on why the proofs of Theorems 1.6 and 1.8 give matching values of ℓ when k is even (thus allowing us to determine the exact quadratic threshold). To make this discussion more convenient, for a k-configuration $\mathcal{C}$ we use $d(\mathcal{C})$ to denote the number of distinct differences in a generic solution to $\mathcal{C}$ (it can be shown that $d(\mathcal{C})$ is well-defined, and that *any* solution to $\mathcal{C}$ has at most $d(\mathcal{C})$ distinct differences).

Let $k = 2s$. Then the idea of our proof of Theorem 1.8 for $t = 1$ is essentially that if $|A - A|$ is small compared to n^2, then the specific k-configuration

$$\mathcal{C}_{\text{sum}} = \{x_1 + x_{s+1} = x_2 + x_{s+2} = \cdots = x_s + x_{2s}\}$$

must occur in A. This gives a lower bound of $g(n, k, \ell) = \Omega(n^2)$ where $\ell = d(\mathcal{C}_{\text{sum}}) + 1$ (because if $|A - A|$ is small, then the occurrence of $\mathcal{C}_{\text{sum}}$ in A has at most $d(\mathcal{C}_{\text{sum}}) < \ell$ distinct differences, contradicting the (k, ℓ)-local property).

On the other hand, when we prove Theorem 1.6 for $c = 2$, we show that there exist n-element sets A with $o(n^2)$ distinct differences such that every k-configuration that occurs in A is 2-good; this gives a bound of $g(n, k, \ell) = o(n^2)$ where ℓ is the minimum value of $d(\mathcal{C})$ over all 2-good k-configurations $\mathcal{C}$.

But as stated in Remark 2.10 (with $k = 2s$ and $d = s + 1$), this minimum is achieved by $\mathcal{C}_{\text{sum}}$ (since it is an equality case of Lemma 2.5, which gives a lower bound on $d(\mathcal{C})$ for 2-good $\mathcal{C}$). So the value of ℓ in our bound of $g(n, k, \ell) = o(n^2)$ actually ends up being $d(\mathcal{C}_{\text{sum}})$; this is why the values of ℓ in the two bounds differ by exactly 1.

In other words, we construct sets A with subquadratically many distinct differences such that one of the 'worst' k-configurations that occurs in our construction — i.e., one with minimal $d(\mathcal{C})$ — actually has to occur in *every* set A with subquadratically many distinct differences; this is what allows us to determine the quadratic threshold.

When k is odd, we will use Proposition 1.7 (in place of Theorem 1.6) to obtain the bound of $o(n^2)$. Meanwhile, to obtain the bound of $\Omega(n^2)$, we extend the bound from Theorem 1.8 to odd k using the following simple observation.

Fact 5.2. *For all $n \geq k \geq 2$ and ℓ, we have $g(n, k, \ell) \geq g(n, k - 1, \ell - k + 1)$.*

Proof. If a set $A \subseteq \mathbb{R}$ satisfies the (k, ℓ)-local property, then we claim that it must also satisfy the $(k - 1, \ell - k + 1)$-local property. To see this, for any $(k - 1)$-element subset $A' \subseteq A$, choose any $a \in A \setminus A'$. Then $A' \cup \{a\}$ has k elements, so by assumption it contains at least ℓ distinct differences. Then removing a removes at most $k - 1$ of these differences, so A' contains at least $\ell - k + 1$ distinct differences, as desired.

Then since $g(n, k, \ell)$ is defined as the *minimum* value of $|A - A|$ over all sets A satisfying the (k, ℓ)-local property, it follows that $g(n, k, \ell) \geq g(n, k - 1, \ell - k + 1)$. $\square$

Now for odd k, we can deduce Corollary 1.2 as well:

- Proposition 1.7 immediately gives that

$$g\left(n, k, \frac{(k + 1)^2}{4} - 4\right) = o(n^2).$$

- Applying Theorem 1.8 with $t = 1$ to $k - 1$ (which is even) and using Fact 5.2 gives that

$$g\left(n, k, \frac{(k + 1)^2}{4}\right) \geq g\left(n, k - 1, \frac{(k + 1)^2}{4} - k + 1\right) = g\left(n, k - 1, \frac{(k - 1)^2}{4} + 1\right) = \Omega(n^2).$$

In particular, these bounds mean that the quadratic threshold for odd k is between $\frac{(k+1)^2}{4} - 3$ and $\frac{(k+1)^2}{4}$.

Remark 5.3. Similarly to Remark 5.1, we now comment on why the values of ℓ in these two bounds differ by 4 (or equivalently, why our bounds on the quadratic threshold for odd k differ by 3). Let $k = 2s + 1$. Then the proof of our bound $g(n, k, \ell) = \Omega(n^2)$ can be viewed as showing that if $|A - A|$ is small compared to n^2, then the k-configuration

$$\mathcal{C}_{\text{sum}} = \{x_1 + x_{s+1} = x_2 + x_{s+2} = \cdots = x_s + x_{2s}\}$$

(in which there is no equation involving x_{2s+1}) must occur in A, and therefore the value of ℓ in the bound is $d(\mathcal{C}_{\text{sum}}) + 1$. (This is not how we phrased the argument, but it is equivalent in some sense, and gives the same value of ℓ.)

On the other hand, we again have a bound of $g(n, k, \ell) = o(n^2)$ where ℓ is the minimum value of $d(\mathcal{C})$ over all 2-good k-configurations $\mathcal{C}$. As stated in Remark 3.6, for $k \geq 7$ this minimum is achieved by

$$\mathcal{C}_{\text{sum}}^+ = \{x_1 + x_{s+1} = x_2 + x_{s+2} = \cdots = x_s + x_{2s}, \, x_1 - x_2 - x_3 + x_{2s+1} = 0\}$$

(since $\mathcal{C}_{\text{sum}}^+$ is an equality case of Lemma 3.1, our bound on $d(\mathcal{C})$ for 2-good $\mathcal{C}$), so the value of ℓ we get in this bound is $d(\mathcal{C}_{\text{sum}}^+)$. The additional equation in $\mathcal{C}_{\text{sum}}^+$ compared to $\mathcal{C}_{\text{sum}}$ means we have $d(\mathcal{C}_{\text{sum}}^+) = d(\mathcal{C}_{\text{sum}}) - 3$, as we gain the equalities $x_{2s+1} - x_2 = x_3 - x_1$, $x_{2s+1} - x_3 = x_2 - x_1$, and $x_{2s+1} - x_{s+1} = x_3 - x_{s+2}$ (or $x_{2s+1} - x_{s+1} = x_2 - x_{s+3}$); this causes the two values of ℓ to differ by 4.

In fact, for $k = 3$ and 5, one can show that the minimum value of $d(\mathcal{C})$ over 2-good k-configurations $\mathcal{C}$ is achieved by $\mathcal{C}_{\text{sum}}$ (for $k = 3$, a valid and AP-free 3-configuration cannot contain any difference equalities; for $k = 5$, by Lemma 3.2 a valid and AP-free 5-configuration contains at most one difference equality), and therefore the quadratic threshold is exactly $d(\mathcal{C}_{\text{sum}}) + 1 = \frac{(k+1)^2}{4}$. However, determining the exact quadratic threshold for odd $k \geq 7$ remains an interesting open problem.

Remark 5.4. It is also natural to ask how $g(n, k, \ell)$ behaves for ℓ immediately below the quadratic threshold. For even k, Theorem 1.6 in fact gives that $g(n, k, \frac{k^2}{4}) = o(n^c)$ for all $c > 2 - \frac{2}{k}$ — for all such c we have $\frac{(c-1)(k-1)}{c} > \frac{k}{2} - 1$, so the value of ℓ for which Theorem 1.6 gives a bound of $o(n^c)$ is still $\frac{k^2}{4}$. On the other hand, the best known lower bound on $g(n, k, \frac{k^2}{4})$ is the bound (1) due to Fish, Pohoata, and Sheffer [5] with $r = 3$, which gives that $g(n, k, \frac{k^2}{4}) = \Omega(n^{3/2-9/k})$ for all $k \geq 18$ divisible by 6.

These two bounds have very different behavior — the exponent $2 - \frac{2}{k}$ in the upper bound grows arbitrarily close to 2 as $k \to \infty$, while the exponent in the lower bound is always at most $\frac{3}{2}$, and is therefore bounded away from 2. It would be interesting to understand which behavior is 'correct' — does there exist a constant $\varepsilon > 0$ not depending on k such that for all k, for all ℓ we either have $g(n, k, \ell) = \Omega(n^2)$ or $g(n, k, \ell) = O(n^{2-\varepsilon})$? (It would also be interesting to answer this question for k restricted to any infinite set, e.g., for all even k.)

5.2. **Intermediate bounds.** For other exponents $1 < c < 2$, the values of ℓ in Theorems 1.6 and 1.8 for which we can bound $g(n, k, \ell)$ above and below by n^c are farther apart. However, both values of ℓ are quadratic in k (considering c and t as constants not depending on k in the two theorems), with leading coefficient tending to 0 as c tends to 1. We will use this fact to deduce Corollaries 1.3 and 1.4. To do so, we first use Fact 5.2 to extend Theorem 1.8 to *all* k, rather than just multiples of 2^t.

Corollary 5.5. *For all positive integers t and k, we have*

$$g\left(n, k, \frac{3^{t-1}k^2}{4^t} + 2^t k + \frac{3^{t-1}+1}{2}\right) = \Omega(n^{1+\frac{1}{2^t-1}}).$$

Proof. Let r be the remainder when k is divided by 2^t, so that $0 \leq r < 2^t$. Then applying Theorem 1.8 to $k - r$ (which is divisible by 2^t) gives

$$(19) \qquad g\left(n, k-r, \frac{3^{t-1}(k-r)^2}{4^t} + \frac{3^{t-1}+1}{2}\right) = \Omega(n^{1+\frac{1}{2^t-1}}).$$

Meanwhile, by repeatedly applying Fact 5.2, we have

$$(20) \qquad g(n, k, \ell) \geq g(n, k-r, \ell - (k-1) - (k-2) - \cdots - (k-r)) \geq g(n, k-r, \ell - kr)$$

for all ℓ. Since $r < 2^t$, we have

$$\frac{3^{t-1}k^2}{4^t} + 2^t k + \frac{3^{t-1}+1}{2} - kr \geq \frac{3^{t-1}k^2}{4^t} + \frac{3^{t-1}+1}{2} \geq \frac{3^{t-1}(k-r)^2}{4^t} + \frac{3^{t-1}+1}{2},$$

so combining (19) and (20) gives the desired bound. (It is possible to improve the lower-order terms in our expression for ℓ by being slightly more careful with these bounds; but only the leading term matters for the following analysis, so we use the above expression for ℓ because it is simpler than the optimized one.) $\qquad \square$

Using this, we can immediately deduce Corollary 1.3, which essentially states that for any given constant c, the threshold for $\Omega(n^c)$ is quadratic in k (the precise statement is slightly stronger than this).

Proof of Corollary 1.3. First we find a constant $a_1 > 0$ such that $g(n, k, a_1 k^2) = o(n^c)$ for large k (providing a lower bound on the threshold for $\Omega(n^c)$ that is quadratic in k).

Choose $a_1 = \frac{1}{2}(\frac{c-1}{c})^2$ (any choice of a_1 with $a_1 < (\frac{c-1}{c})^2$ would work for the same reason; we choose one for concreteness). Then for all sufficiently large k we have

$$a_1 k^2 \leq \left\lceil \frac{(c-1)(k-1)}{c} \right\rceil^2,$$

so by Theorem 1.6 we have

$$g(n, k, a_1 k^2) \leq g\left(n, k, \left\lceil \frac{(c-1)(k-1)}{c} \right\rceil^2\right) = o(n^c).$$

Now we find a constant $0 < a_2 < \frac{1}{2}$ such that $g(n, k, a_2 k^2) = \Omega(n^c)$ for large k (providing an upper bound for the threshold for $\Omega(n^c)$ that is quadratic in k); note that this automatically implies $a_2 > a_1$.

Let t be the largest positive integer such that $c \leq 1 + \frac{1}{2^t - 1}$, and choose $a_2 = \frac{1}{2}(\frac{3}{4})^{t+1}$ (any $a_2 > \frac{3^{t-1}}{4^t}$ would work for the same reason; we choose this one for concreteness and because it makes the computations in Remark 5.6 slightly nicer). Then for all sufficiently large k we have

$$a_2 k^2 \geq \frac{3^{t-1} k^2}{4} + 2^t k + \frac{3^{t-1} + 1}{2},$$

so by Corollary 5.5 we have

$$g(n, k, a_2 k^2) \geq g\left(n, k, \frac{3^{t-1} k^2}{4} + 2^t k + \frac{3^{t-1} + 1}{2}\right) = \Omega(n^c).$$

Finally, note that as $c \to 1$, we have $t \to \infty$ and therefore $a_2 = \frac{1}{2}(\frac{3}{4})^{t+1} \to 0$; since $a_1 < a_2$, this means we have $a_1 \to 0$ as well. $\qquad\square$

Remark 5.6. We now analyze the discrepancy between the values of a_1 and a_2 given in the above proof. To do so, we first obtain bounds on a_2 purely in terms of c — note that $c \leq 1 + \frac{1}{2^s - 1}$ if and only if $2^s \leq \frac{c}{c-1}$, and therefore the value of t in the above proof must satisfy

$$2^t \leq \frac{c}{c-1} < 2^{t+1}.$$

Since we chose $a_2 = \frac{1}{2}(\frac{3}{4})^{t+1}$, this means we have

$$\frac{3}{4}\left(\frac{c-1}{c}\right)^{\log_2 \frac{4}{3}} \leq 2a_2 < \left(\frac{c-1}{c}\right)^{\log_2 \frac{4}{3}}.$$

Meanwhile, we set $a_1 = \frac{1}{2}(\frac{c-1}{c})^2$, so we have

$$\frac{3}{4}(2a_1)^{\log_4 \frac{4}{3}} \leq 2a_2 < (2a_1)^{\log_4 \frac{4}{3}}.$$

We can also prove Corollary 1.4, which gives upper and lower bounds of the form n^c when ℓ is quadratic in k, in the same way.

Proof of Corollary 1.4. To obtain the lower bound of $g(n, k, ak^2) = \Omega(n^{c_1})$, let t be the smallest positive integer such that $a > \frac{3^{t-1}}{4^t}$, and let $c_1 = 1 + \frac{1}{2^t - 1}$. Then for all sufficiently large k we have

$$ak^2 \geq \frac{3^{t-1}}{4^t} k^2 + 2^t k + \frac{3^{t-1} + 1}{2},$$

so by Corollary 5.5 we have

$$g(n, k, ak^2) \geq g\left(n, k, \frac{3^{t-1}}{4^t} k^2 + 2^t k + \frac{3^{t-1} + 1}{2}\right) = \Omega(n^{c_1}).$$

We now turn to the upper bound. Given $0 < a < \frac{1}{4}$, choose $1 < c_2 < 2$ such that $a < (\frac{c_2-1}{c_2})^2$ (this is possible because $(\frac{c-1}{c})^2$ ranges over $(0, \frac{1}{4})$ as c ranges over $(1, 2)$); it is possible to choose such a c_2 as an explicit function of a such that $c_2 \to 0$ as $a \to 0$. Then for all sufficiently large k we have

$$ak^2 \leq \left\lceil \frac{(c_2 - 1)(k-1)}{c_2} \right\rceil^2,$$

so by Theorem 1.6, we have

$$g(n, k, ak^2) \leq g\left(n, k, \left\lceil \frac{(c_2 - 1)(k - 1)}{c_2} \right\rceil^2\right) = o(n^{c_2}).$$

(Note that these two bounds automatically imply that $c_1 < c_2$, so $c_1 \to 1$ as $a \to 0$ as well.) $\qquad\square$

5.3. **On the number of possible exponents.** As stated in Subsection 1.4, for any fixed k, we can attempt to understand which exponents 'appear' in our asymptotics for $g(n, k, \ell)$ over all ℓ; to formalize this, we define the set

$$S_k = \left\{ \liminf_{n \to \infty} \frac{\log g(n, k, \ell)}{\log n} \mid 0 \leq \ell \leq \binom{k}{2} \right\}.$$

Clearly $1 \in S_k$ (as $g(n, k, k - 1) = n - 1$), and $2 \in S_k$ for $k \geq 4$ (as $g(n, k, \binom{k}{2}) = \binom{n}{2}$). The currently known bounds on $g(n, k, \ell)$ are not tight enough to allow us to determine any other values that must be in S_k. However, we *can* use Theorems 1.6 and 1.8 to show that $|S_k|$ tends to infinity as k does with a growth rate of at least $\log \log k$, as stated in Corollary 1.5.

Proof of Corollary 1.5. First, it suffices to prove there exists a such that $|S_k| \geq a \log \log k$ for all sufficiently large k; then we can adjust a to deal with the finitely many remaining values of k (as $|S_k| \geq 1$ for all k).

For notational convenience, let

$$e(\ell) = \liminf_{n \to \infty} \frac{\log g(n, k, \ell)}{\log n}.$$

Since $g(n, k, \ell)$ is (weakly) increasing in ℓ, the same is true of $e(\ell)$. We wish to find at least $a \log \log k$ values of ℓ with distinct values of $e(\ell)$ (for some absolute constant $a > 0$). The idea of the proof is to choose values $\ell_1 > \ell_2 > \cdots$ sufficiently spaced out so that for each i, the lower bound on $e(\ell_i)$ given by Theorem 1.8 (or rather, Corollary 5.5) is greater than the upper bound on $e(\ell_{i+1})$ given by Theorem 1.6, which implies that we must have $e(\ell_i) > e(\ell_{i+1})$.

To do so, for each positive integer t, let

$$\ell^+(t) = \frac{3^{t-1}k^2}{4^t} + 2^t k + \frac{3^{t-1} + 1}{2}$$

(this is the value of ℓ for which Corollary 5.5 gives $g(n, k, \ell) = \Omega(n^{1 + \frac{1}{2^t - 1}})$), and let

$$\ell^-(t) = \left\lceil \frac{k - 1}{2^t} \right\rceil^2$$

(this is the value of ℓ for which Theorem 1.6 gives $g(n, k, \ell) = o(n^{1 + \frac{1}{2^t - 1}})$). Then we have

$$e(\ell^-(t)) \leq 1 + \frac{1}{2^t - 1} \leq e(\ell^+(t))$$

for all positive integers t.

We will choose $\ell_i = \ell^+(t_i)$ for certain values of t_i. This will automatically give lower bounds on our values of $e(\ell_i)$; the following claim will allow us to obtain comparable upper bounds.

Claim 5.7. *For any integer $t \geq 2$ such that $k \geq 3 \cdot 4^{10t}$, we have $\ell^+(10t) \leq \ell^-(t + 1)$.*

Proof. First, since $k - 1 \geq \frac{k}{2}$, we have

$$\ell^-(t + 1) = \left\lceil \frac{k - 1}{2^{t+1}} \right\rceil^2 \geq \left(\frac{k}{2^{t+2}} \right)^2 = \frac{1}{4^{t+2}} k^2.$$

Meanwhile, let $s = 10t$, so that

$$\ell^+(s) = \frac{3^{s-1}}{4^s} k^2 + 2^s k + \frac{3^{s-1} + 1}{2} \leq \frac{3^{s-1}}{4^s} k^2 + 2^s k + 3^{s-1}.$$

The condition $k \geq 3 \cdot 4^s$ guarantees that the first term is the largest of the three terms (the ratio of the first and second terms is $\frac{3^{s-1}}{8^s} \cdot k \geq \frac{1}{3} \cdot \left(\frac{3}{8} \right)^s \cdot 3 \cdot 4^s \geq 1$, and the ratio of the first and third is $\frac{k^2}{4^s} \geq 1$), so

$$\ell^+(s) \leq 3 \cdot \frac{3^{s-1}}{4^s} k^2 = \left(\frac{3}{4} \right)^s k^2.$$

Finally, for all $t \geq 2$, we have

$$\frac{1}{4^{t+2}} \geq \frac{1}{4^{2t}} \geq \left(\frac{3}{4}\right)^{10t}$$

(using the fact that $\frac{1}{4} \geq \left(\frac{3}{4}\right)^5$). Combining these bounds gives $\ell^+(10t) \leq \ell^-(t+1)$, as desired. $\qquad\square$

Now let $r = \left\lfloor \log_{10} \log_4 \frac{k}{3} \right\rfloor$. For each integer $1 \leq i \leq r$, define $t_i = 10^i$, and let $\ell_i = \ell^+(t_i)$. Then for all $1 \leq i \leq r$ we have $k \geq 3 \cdot 4^{t_i}$, so Claim 5.7 implies that for all $1 \leq i \leq r-1$ we have

$$\ell^+(t_{i+1}) \leq \ell^-(t_i + 1),$$

and therefore

$$e(\ell^+(t_{i+1})) \leq e(\ell^-(t_i+1)) \leq 1 + \frac{1}{2^{t_i+1}-1} < 1 + \frac{1}{2^{t_i}-1} \leq e(\ell^+(t_i)).$$

So $e(\ell_r) < e(\ell_{r-1}) < \cdots < e(\ell_1)$, which means the values of $e(\ell_i)$ for $1 \leq i \leq r$ are all distinct. (Note that for sufficiently large k, all our values of ℓ_i are valid, i.e., $0 \leq \ell^+(t_i) \leq \binom{k}{2}$ for all our values of t_i. To see this, clearly $\ell^+(t) \geq 0$ for all t. Meanwhile for sufficiently large k we have $\ell_1 = \ell^+(10) \leq \binom{k}{2}$ (as the leading coefficient in our expression for $\ell^+(10)$ is $\frac{3^9}{4^{10}} < \frac{1}{2}$), and the above argument in particular implies that $\ell_{i+1} < \ell_i$ for all $1 \leq i \leq r-1$, so $\ell_i \leq \binom{k}{2}$ for all $1 \leq i \leq r$.)

This gives us r values $0 \leq \ell \leq \binom{k}{2}$ with distinct values of $e(\ell)$, so

$$|S_k| \geq r \geq \log_{10} \log_4 \frac{k}{3} - 1.$$

Letting $a = \frac{1}{2\log 10}$, we then have $|S_k| \geq a \log \log k$ for all sufficiently large k, as desired. $\qquad\square$

References

[1] F. A. Behrend. On sets of integers which contain no three terms in arithmetical progression. *Proceedings of the National Academy of Sciences*, 32(12):331–332, 1946.

[2] David Conlon, Jacob Fox, Choongbum Lee, and Benny Sudakov. The Erdős–Gyárfás problem on generalized Ramsey numbers. *Proceedings of the London Mathematical Society*, 110(1):1–18, 2015.

[3] Paul Erdős. On some metric and combinatorial geometric problems. *Discrete Mathematics*, 60:147–153, 1986.

[4] Paul Erdős and András Gyárfás. A variant of the classical Ramsey problem. *Combinatorica*, 17:459–467, 1997.

[5] Sara Fish, Ben Lund, and Adam Sheffer. A construction for difference sets with local properties. *European Journal of Combinatorics*, 79:237–243, 2019.

[6] Sara Fish, Cosmin Pohoata, and Adam Sheffer. Local properties via color energy graphs and forbidden configurations. *SIAM Journal on Discrete Mathematics*, 34(1):177–187, 2020.

[7] Anqi Li. Progress on local properties problems of difference sets. *European Journal of Combinatorics*, 108(103618), 2022.

[8] Cosmin Pohoata and Adam Sheffer. Local properties in colored graphs, distinct distances, and difference sets. *Combinatorica*, 39(3):705–714, 2019.

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

Email address: sanjanad@mit.edu