

Testing linear-invariant properties

Jonathan Tidor and Yufei Zhao
 Department of Mathematics
 Massachusetts Institute of Technology
 Cambridge, MA 02139, USA
 {jtidor,yufeiz}@mit.edu

Abstract—Fix a prime p and a positive integer R . We study the property testing of functions $\mathbb{F}_p^n \rightarrow [R]$. We say that a property is testable if there exists an oblivious tester for this property with one-sided error and constant query complexity. Furthermore, a property is proximity oblivious-testable (PO-testable) if the test is also independent of the proximity parameter ϵ . It is known that a number of natural properties such as linearity and being a low degree polynomial are PO-testable. These properties are examples of linear-invariant properties, meaning that they are preserved under linear automorphisms of the domain. Following work of Kaufman and Sudan, the study of linear-invariant properties has been an important problem in arithmetic property testing.

A central conjecture in this field, proposed by Bhattacharyya, Grigorescu, and Shapira, is that a linear-invariant property is testable if and only if it is semi subspace-hereditary. We prove two results, the first resolves this conjecture and the second classifies PO-testable properties.

- 1) A linear-invariant property is testable if and only if it is semi subspace-hereditary.
- 2) A linear-invariant property is PO-testable if and only if it is locally characterized.

Our innovations are two-fold. We give a more powerful version of the compactness argument first introduced by Alon and Shapira. This relies on a new strong arithmetic regularity lemma in which one mixes different levels of Gowers uniformity. This allows us to extend the work of Bhattacharyya, Fischer, Hatami, Hatami, and Lovett by removing the bounded complexity restriction in their work. Our second innovation is a novel recoloring technique called patching. This Ramsey-theoretic technique is critical for working in the linear-invariant setting and allows us to remove the translation-invariant restriction present in previous work.

Index Terms—property testing; sublinear time algorithms; removal lemmas; higher-order Fourier analysis;

I. INTRODUCTION

In property testing, the aim is to find randomized algorithms that distinguish objects that have some property from those that are far from satisfying the property by querying the given large object at a small number of locations. Property testing emerged from the linearity test of Blum, Luby, and Rubinfeld [10], and was formally defined and systematically studied by Rubinfeld and Sudan [29] and Goldreich, Goldwasser, and Ron [14]. There have been important developments especially in

the following two settings: graph property testing and arithmetic property testing.

Two representative problems are: (1) given a large graph, test whether the graph is triangle-free or ϵ -far from triangle-free (an n -vertex graph is ϵ -far from a graph property if one needs to add and/or remove more than ϵn^2 edges in order to satisfy the property), and (2) given a function $f: \mathbb{F}_p^n \rightarrow \mathbb{F}_p$, test whether f is linear or ϵ -far from linear (a function is ϵ -far from an arithmetic property if one needs to change the value of the function on more than an ϵ -fraction of the domain in order to satisfy the property). In both cases, it is known that one can achieve the desired goal by sampling a fixed number of entries repeatedly $C(\epsilon)$ times. For testing whether a graph is triangle-free [30], one samples a uniform random triple of vertices and checks whether they form a triangle, and for testing linearity [10], one samples $x, y \in \mathbb{F}_p^n$ uniformly and checks if $f(x) + f(y) = f(x + y)$.

In this paper we give a property testing algorithm for a very general class of arithmetic properties. The goal is to determine whether a function $f: \mathbb{F}_p^n \rightarrow [R] := \{1, \dots, R\}$ (with fixed prime p and positive integer R) satisfies some given property or is ϵ -far from satisfying the property. All the properties we consider are *linear-invariant* in the sense that they are invariant under automorphisms of the vector space \mathbb{F}_p^n . Linear-invariant properties form an important general class of arithmetic properties, e.g., the work of Kaufman and Sudan [26] “highlights linear-invariance as a central theme in algebraic property testing.”

We say that a property \mathcal{P} is *testable* if there exists an oblivious tester with one-sided error (and constant query complexity) for the property. A tester for \mathcal{P} produces a positive integer $d = d(\epsilon)$ and an oracle provides the tester with the restriction $f|_U$ where U is a uniform random d -dimensional linear subspace of the domain (if the domain is large enough that such a subspace exists; if the domain has dimension strictly less than d , the oracle provides the tester with all of f). We require our tester accepts functions f satisfying \mathcal{P} with probability 1 and reject functions that are ϵ -far from satisfying \mathcal{P} with probability at least $\delta = \delta(\epsilon)$ for some function $\delta: (0, 1) \rightarrow (0, 1)$. Furthermore, we say that \mathcal{P}

is *proximity oblivious-testable* (PO-testable) if $d = d(\epsilon)$ is a constant independent of ϵ . The idea of PO-testability was introduced by Goldreich and Ron [15] who, among other results, classified the PO-testable graph properties.

One surprising feature of property testing is that many natural properties, such as linearity, are testable and even PO-testable. A key feature of linearity is that it is subspace-hereditary meaning that if $f: \mathbb{F}_p^n \rightarrow \mathbb{F}_p$ is linear, then the same is true for $f|_U$ for every linear subspace $U \leq \mathbb{F}_p^n$. To be precise, we say that a linear-invariant property \mathcal{P} is *subspace-hereditary* if for every $f: \mathbb{F}_p^n \rightarrow [R]$ satisfying \mathcal{P} and every linear subspace $U \leq \mathbb{F}_p^n$, the restriction $f|_U$ also satisfies \mathcal{P} .

A central conjecture in this field, first proposed by Bhattacharyya, Grigorescu, and Shapira, is that all linear-invariant, subspace-hereditary properties are testable [9, Conjecture 4]. In fact, they conjecture that the slightly larger class of *semi subspace-hereditary* properties are testable and prove that no other properties can be tested.

Definition I.1. A linear-invariant property \mathcal{P} is *semi subspace-hereditary* if there exists a subspace-hereditary property \mathcal{Q} such that

- (i) every function satisfying \mathcal{P} also satisfies \mathcal{Q} ;
- (ii) for all $\epsilon > 0$, there exists $N(\epsilon)$ such that if $f: \mathbb{F}_p^n \rightarrow [R]$ satisfies \mathcal{Q} and is ϵ -far from satisfying \mathcal{P} , then $n < N(\epsilon)$.

It is known that there are subspace-hereditary properties where the dimension d sampled must grow as the proximity parameter ϵ approaches 0. To be PO-testable, a property must satisfy the following more restrictive condition.

Definition I.2. A linear-invariant property \mathcal{P} is *locally characterized* if there exists some d such that the following holds. For every $f: \mathbb{F}_p^n \rightarrow [R]$ with $n \geq d$, the function f satisfies \mathcal{P} if and only if $f|_U$ satisfies \mathcal{P} for every $U \leq \mathbb{F}_p^n$ of dimension d .

Our first result is a resolution of the conjecture of Bhattacharyya, Grigorescu, and Shapira, classifying the testable linear-invariant properties. Our second result is a classification of the PO-testable linear-invariant properties.

Theorem I.3. *A linear-invariant property is testable if and only if it is semi subspace-hereditary.*

Theorem I.4. *A linear-invariant property is PO-testable if and only if it is locally characterized.*

Remark I.5. Note that under our definition, the tester does not know the dimension of the domain. This rules out some “unnatural” properties such as those properties that behave differently depending on whether the dimension of the domain is even or odd.

Previous work in arithmetic property testing has focused on a number of special cases including monotone properties [27], [31], “complexity 1” properties over \mathbb{F}_2 [9], and bounded complexity translation-invariant properties [7].

We note that very little was previously known about general linear-invariant properties. One simple way to define a class of linear-invariant properties can be done by, for example, choosing an arbitrary subset of “allowable” maps $\mathbb{F}_p^2 \rightarrow [R]$ and defining a property of functions $\mathbb{F}_p^n \rightarrow [R]$ to consist of those whose restriction to every 2-dimensional linear subspace is allowable. Even this class of 2-dimensionally-defined patterns was not known to be testable in general prior to this work.

Our innovations are two-fold. We prove a strong arithmetic regularity lemma which, unlike previous arithmetic regularity lemmas, mixes different levels of Gowers uniformity. This allows us to give a more powerful version of the *compactness argument* first introduced by Alon and Shapira [4]. With this tool we can remove the *bounded complexity* restriction that was present in all previous work.

Our second innovation is a novel recoloring technique we call *patching*. This technique is critical for working in the linear-invariant setting and allows us to handle an important obstacle encountered by previous works.

In the rest of this section we give a summary of the proof of the main theorem and its relation to previous work. This proceeding version omits many proofs and some intermediate technical results. All omitted details can be found in the full version of the paper [34].

A. Graph removal lemmas and property testing

We begin with an overview of graph removal lemmas and their proof techniques (see the survey [11] for a more detailed discussion).

The foundational result in this field is the triangle removal lemma of Ruzsa and Szemerédi [30]. This result states that for all $\epsilon > 0$ there exists $\delta > 0$ such that any n -vertex graph with at most δn^3 triangles can be made triangle-free by removing ϵn^2 edges. It is immediate from the definitions that this result implies that triangle-freeness is testable (and in fact PO-testable). In general we will work with such *removal lemmas* and deduce corresponding property testing results from them. The triangle removal lemma was generalized to the graph removal lemma, first stated explicitly by Alon, Duke, Lefmann, Rödl, and Yuster [1] and by Füredi [13].

A key tool for proving the graph removal lemma is a regularity lemma, namely Szemerédi’s graph regularity lemma. Roughly speaking, the proof proceeds by using this regularity lemma to partition the input graph G into a small number of structured components. Then we “clean up” G by removing at most ϵn^2 edges. This is done in

such a way that either the resulting graph is H -free or the original graph G contains many copies of H .

An important extension of the graph removal lemma is the induced graph removal lemma, proved by Alon, Fischer, Krivelevich, and Szegedy [2]. The induced graph removal lemma states that for every graph H , for all $\epsilon > 0$ there exists $\delta > 0$ such that every n -vertex graph with at most $\delta n^{v(H)}$ induced copies of H can be made induced H -free by adding and/or removing at most ϵn^2 edges (here *induced H -free* means not containing any induced subgraph isomorphic to H).

The original proof of the induced graph removal lemma relies on an extension of Szemerédi’s graph regularity lemma known as the “strong regularity lemma.” Using such a regularity lemma combined with a random sampling argument, one can produce a “regular model”, that is, a large induced subgraph $X := G[U]$ (on a constant fraction of the vertices of G) that is very regular and approximates the original graph well in a certain sense. Then we “clean up” G by adding and/or removing at most ϵn^2 edges in such a way that if the resulting graph is not induced H -free then X (in the original graph) must contain many induced copies of H .

Both the graph removal lemma and the induced graph removal lemma can be easily extended to remove any finite collection \mathcal{H} of graphs. Alon and Shapira [4] extended the induced graph removal lemma to an infinite collection of graphs. Namely they prove that for a (possibly infinite) set \mathcal{H} of graphs and for $\epsilon > 0$ there exist $\delta > 0$ and k such that the following holds: if G is an n -vertex graph with at most $\delta n^{v(H)}$ copies of H for all $H \in \mathcal{H}$ with k or fewer vertices, then G can be made induced \mathcal{H} -free by adding and/or removing at most ϵn^2 edges (meaning the modified graph has no induced subgraph isomorphic to H for every $H \in \mathcal{H}$). Despite the strange statement of this result (the hypothesis that G has low H -density only for the small graphs $H \in \mathcal{H}$) this theorem immediately implies (and is equivalent to) the fact that every hereditary graph property is testable with constant query-complexity and one-sided error.

This series of works, in addition to being important results in their own right, gives a framework for proving constant query-complexity property testing algorithms in other settings, given an appropriate regularity lemma. For example a hypergraph regularity lemma is known, proven by Gowers [17] and independently by Rödl et al. [28]. Using the above framework, one can use the hypergraph regularity lemma to deduce an infinite induced hypergraph removal lemma [29]. Consequently, every hereditary hypergraph property is testable with constant query-complexity and one-sided error. We will be interested in seeing if this framework, combined with an appropriate arithmetic regularity lemma, can prove an infinite induced arithmetic removal lemma.

B. Arithmetic analogs

The problem of property testing for functions $f: \mathbb{F}_p^n \rightarrow [R]$ has been intensively studied, starting with the classic work of Blum, Luby, and Rubinfeld [10] on linearity testing. Much of the work focuses on testing whether some function $f: \mathbb{F}_p^n \rightarrow \mathbb{F}_p$ has certain algebraic properties (e.g., a polynomial of some given type) [3], [26]. There is also much interest in testing properties that do not arrive from algebraic characterizations. Below we give an overview of the developments related to property testing in \mathbb{F}_p^n from a perspective that is parallel to the graph regularity method developments discussed earlier.

The first arithmetic regularity lemma was proved by Green [19] using Fourier-analytic techniques, and it laid the groundwork for further developments of the regularity method in the arithmetic setting. These regularity lemmas have since found many applications in additive combinatorics and related fields. In particular, combined with the graph removal framework described above, Green’s regularity lemma is suitable for proving an arithmetic removal lemma for “complexity 1” systems of linear forms (see Section III for the definition of complexity; roughly speaking, a system of linear forms is complexity 1 if it can be controlled by Fourier-analytic means); e.g., see [6].

Král’, Sera, and Vena [27] and independently Shapira [31] bypass the need for an arithmetic regularity lemma and prove the full arithmetic removal lemma by a direct reduction from the hypergraph removal lemma. Their results imply that all linear-invariant, subspace-hereditary *monotone* properties are testable with constant query-complexity and one-sided error. (A property of functions $\mathbb{F}_p^n \rightarrow \{0, 1\}$ is *monotone* if changing 1’s to 0’s preserves the property.)

Note that the above result is an arithmetic removal lemma and not an *induced* arithmetic removal lemma (hence the restriction to monotone properties). Due to the nature of the reduction, the techniques do not seem to be capable of deducing the induced arithmetic removal lemma from the induced hypergraph removal lemma.

An alternative approach is to apply the strong graph regularity approach [2] of proving the induced graph removal lemma to Green’s arithmetic regularity lemma. However there is also a major obstacle to the approach, related to the fact that the origin plays a special role in a vector space while there is no corresponding feature of graphs. It turns out that it is not always possible to regularize the space in a neighborhood of the origin [20].

Bhattacharyya, Grigorescu, and Shapira [9] managed to overcome this obstacle in the special case of vector spaces over \mathbb{F}_2 . They follow the above strategy, implementing the strong regularity idea [2] in the style of Green’s arithmetic regularity [19] along with one additional tool, namely a Ramsey-theoretic result, to

prove an infinite induced arithmetic removal lemma for complexity 1 patterns over \mathbb{F}_2 . Unfortunately, it is known [20] that this Ramsey-theoretic result fails over all finite fields other than \mathbb{F}_2 .

Bhattacharyya, Fischer, and Lovett [8] managed to overcome this obstacle in a different special case, namely for translation-invariant patterns. When all patterns considered are translation-invariant, the origin no longer plays a special role and one can essentially ignore it while carrying out the strong regularity framework. In addition, [8] allows one to handle higher complexity patterns, which requires developing and applying tools from higher-order Fourier analysis.

Higher-order Fourier analysis plays a central role in modern additive combinatorics. These techniques were initiated by Gowers [16] in his celebrated new proof of Szemerédi’s theorem, and further developed in a sequence of works by Green, Tao, and Ziegler [21], [22], [23] settling classical conjectures on the asymptotics of prime numbers patterns. A parallel theory of higher-order Fourier analysis was developed in finite field vector spaces by Bergelson, Tao, and Ziegler [5], [32], [33], leading to an inverse Gowers theorem over finite field vector spaces.

For applications to property testing, this line of work culminated in the work of Bhattacharyya, Fischer, Hatami, Hatami, and Lovett [7] (extending [8]), who applied the inverse Gowers theorem over finite fields and developed further equidistribution tools to prove an infinite induced arithmetic removal lemma for all linear-invariant, subspace-hereditary properties that are also translation-invariant and bounded-complexity. Their work follows the strong regularity framework of [2], [4]. Our results improve upon this work by removing the translation-invariant and bounded-complexity restrictions.

In addition to their property testing algorithm, Bhattacharyya, Fischer, Hatami, Hatami, and Lovett [7] proved that a large class of somewhat algebraically structured properties are indeed affine-invariant, subspace-hereditary, and locally characterized. These are the so-called “degree-structural properties”. A simple extension of their result [25, Theorem 16.3] implies that the larger class of “homogeneous degree-structural properties” are linear-invariant, linear subspace-hereditary (but not affine-invariant and not subspace-hereditary), and locally characterized, and thus these properties are testable by our main theorem. As an example, one can test whether a function $\mathbb{F}_p^n \rightarrow \mathbb{F}_p$ can be written as $A^2 + B^2$ where both A and B are homogeneous polynomials of some given degree d .

C. Our contributions

1) *Patching*: In this paper, building on the authors’ earlier work with Fox [12] for complexity 1 patterns, we develop a new technique called “patching” that allows us to overcome the obstacle faced by earlier approaches, namely that a neighborhood of the origin cannot be regularized and fails certain Ramsey properties (unless working over \mathbb{F}_2). In essence, the patching result states that if there exists some map $f: \mathbb{F}_p^n \rightarrow [R]$ that has low density of some colored patterns \mathcal{H} for n large enough, then for all m there must exist some map $g: \mathbb{F}_p^m \rightarrow [R]$ that has no \mathcal{H} -instances.

Theorem I.6 (Informal patching result). *For every set of colored patterns \mathcal{H} , there exist $\epsilon_0 > 0$ and n_0 such that the following holds. Either:*

- *for every n , there exists a function $f: \mathbb{F}_p^n \rightarrow [R]$ that is \mathcal{H} -free; or*
- *for every function $f: \mathbb{F}_p^n \rightarrow [R]$ with $n \geq n_0$, the H -density in f is at least ϵ_0 for some $H \in \mathcal{H}$.*

Our proof proceeds in two steps. First, as in [7], following the strong regularity framework of [2] for proving induced graph removal lemmas, we apply a strong arithmetic regularity lemma, which produces a partition \mathfrak{B} of \mathbb{F}_p^n and a “regular model” $X \subseteq \mathbb{F}_p^n$ made up of a randomly sampled set of atoms from \mathfrak{B} . Unlike in the graph setting, we cannot ensure that the map $f: \mathbb{F}_p^n \rightarrow [R]$ is very regular on every atom of $\mathfrak{B}|_X$. In particular, it may be impossible to guarantee that f is regular on the atom containing the origin. Instead we only ensure that almost every atom of X is very regular. Unlike earlier proofs of removal lemmas, our “recoloring algorithm” has two components: for the regular atoms we “clean up” f as usual, while for the irregular atoms we apply our patching result. Our patching result implies that there is some new global coloring $g: \mathbb{F}_p^n \rightarrow [R]$ that avoids some appropriate set of colored patterns. To complete the proof we “patch” f by replacing it by g on all of the irregular atoms. If f has low density of some set of colored pattern, then our argument shows that these pattern cannot appear in the recoloring, thereby completing the proof of the induced arithmetic removal lemma.

Our proof does not give effective bounds on the rejection probability function $\delta(\epsilon)$ guaranteed by Theorem I.3. The ineffectiveness is due to the fact that the current best-known bounds on the inverse theorem for non-classical polynomials are ineffective (the same occurs in [7]).

2) *Unbounded complexity*: The technique used to handle infinite removal lemmas is a compactness argument first introduced by Alon and Shapira [4] in the graph setting. A key ingredient of their proof is a strong regularity lemma.

Bhattacharyya, Fischer, Hatami, Hatami, and Lovett [7] prove that all linear-invariant subspace-hereditary properties that are also translation-invariant and bounded-complexity are testable. Their result follows from an infinite removal lemma for arithmetic patterns of bounded complexity. The proof of this result involves a strong arithmetic regularity lemma and a compactness argument in the spirit of Alon and Shapira.

To remove the bounded complexity assumption from [7], we prove a new strong arithmetic regularity lemma obtained by iterating a weaker arithmetic regularity lemma. The key innovation here is the level of Gowers uniformity used in each iteration is allowed to increase at each step of the process.

II. COLORED PATTERNS AND REMOVAL LEMMAS

Theorem I.3 and Theorem I.4 both follow from an arithmetic removal lemma for colored linear patterns. In this section we define these objects and state the main removal lemma.

Definition II.1. A **linear form** over \mathbb{F}_p in ℓ variables is an expression L of the form

$$L(x_1, \dots, x_\ell) = \sum_{i=1}^{\ell} c_i x_i$$

with $c_i \in \mathbb{F}_p$. For any \mathbb{F}_p -vector space V , the linear form L gives rise to a function $L: V^\ell \rightarrow V$ that is linear in each variable.

Definition II.2. For a prime p and a finite set \mathcal{S} , an **\mathcal{S} -colored pattern** over \mathbb{F}_p consisting of m linear forms in ℓ variables is a pair (\mathbf{L}, ψ) given by a system $\mathbf{L} = (L_1, \dots, L_m)$ of m linear forms in ℓ variables and a coloring $\psi: [m] \rightarrow \mathcal{S}$. Given a finite-dimensional \mathbb{F}_p -vector space V and a function $f: V \rightarrow \mathcal{S}$, an (\mathbf{L}, ψ) -**instance** in f is some $\mathbf{x} \in V^\ell$ such that $f(L_i(\mathbf{x})) = \psi(i)$ for all $i \in [m]$. An instance is called **generic** if x_1, \dots, x_ℓ are linearly independent. We say that (\mathbf{L}, ψ) is **translation-invariant** if the coefficient of x_1 is 1 in each of L_1, \dots, L_m .

Given a finite-dimensional \mathbb{F}_p -vector space V and functions $f_1, \dots, f_m: V \rightarrow [-1, 1]$, we write

$$\Lambda_{\mathbf{L}}(f_1, \dots, f_m) := \mathbb{E}_{\mathbf{x} \in V^k} [f_1(L_1(\mathbf{x})) \cdots f_m(L_m(\mathbf{x}))].$$

Definition II.3. For an \mathcal{S} -colored pattern over \mathbb{F}_p consisting of m linear forms in k variables (\mathbf{L}, ψ) , a finite dimensional \mathbb{F}_p -vector space V , and a function $f: V \rightarrow \mathcal{S}$, define the (\mathbf{L}, ψ) -**density** in f to be $\Lambda_{\mathbf{L}}(f_1, \dots, f_m)$ where $f_i := 1_{f^{-1}(\psi(i))}$ for each $i \in [m]$.

Our main removal lemma is the following result.

Theorem II.4 (Main removal lemma). *Fix a prime p and a finite set \mathcal{S} . Let \mathcal{H} be a (possibly infinite) set of*

\mathcal{S} -colored patterns over \mathbb{F}_p . For every $\epsilon > 0$, there exists a finite set $\mathcal{H}_\epsilon \subseteq \mathcal{H}$ and $\delta = \delta(\epsilon, \mathcal{H}) > 0$ such that the following holds. Let V be a finite-dimensional \mathbb{F}_p -vector space. If $f: V \rightarrow \mathcal{S}$ has H -density at most δ for every $H \in \mathcal{H}_\epsilon$, then there exists a recoloring $g: V \rightarrow \mathcal{S}$ that agrees with f on all but an at most ϵ -fraction of V such that g has no generic H -instances for every $H \in \mathcal{H}$.

There are several difficulties in the proof of the main removal lemma. The first is that individual patterns $H \in \mathcal{H}$ may have “infinite complexity”. Second, the set of patterns \mathcal{H} may be infinite. Complicating this, even if all patterns in \mathcal{H} have finite complexity, these complexities can be unbounded. Finally, there are major difficulties related to the fact that the patterns in \mathcal{H} are not necessarily translation-invariant.

We use a trick called “projectivization” to reduce to the case where all patterns have finite complexity. To do this, we need a slightly modified version of the main removal lemma that we call the projective removal lemma (Theorem II.8).

A “compactness argument” due to Alon and Shapira [4] reduces the problem of an infinite collection of patterns to a finite one at the expense of requiring a stronger arithmetic regularity lemma. If the collection of patterns is all complexity at most d , we only require a strong U^{d+1} -regularity lemma with rapidly decreasing error parameter. In the most general case when the collection of patterns has unbounded complexity we require an even stronger regularity lemma where the error parameter rapidly decreases and the degree of the uniformity norm rapidly increases.

Unless we restrict to the special case where all patterns in \mathcal{H} are translation-invariant, the origin of the vector space plays a special role. This is unfortunate because it is impossible to regularize a function in the neighborhood of the origin. Since regularity methods are useless here, we turn to a new technique called patching, originally introduced by the authors and Fox [12], to deal with the portions of the vector space that cannot be regularized.

Definition II.5. Let \mathcal{S} be a finite set equipped with a group action of \mathbb{F}_p^\times that we denote $c \cdot s$ for $c \in \mathbb{F}_p^\times$ and $s \in \mathcal{S}$. Given a finite-dimensional \mathbb{F}_p -vector space V , a function $f: V \rightarrow \mathcal{S}$ is **projective** if it preserves the action of \mathbb{F}_p^\times , i.e., $f(cx) = c \cdot f(x)$ for all $c \in \mathbb{F}_p^\times$ and all $x \in V$.

Definition II.6. A list of linear forms $\mathbf{L} = (L_1, \dots, L_m)$ is **finite complexity** if no form is identically equal to zero, i.e., $L_i \not\equiv 0$ for all $i \in [m]$, and no two forms are linearly dependent, i.e., $L_i \not\equiv cL_j$ for all $i \neq j$ and $c \in \mathbb{F}_p$.

Definition II.7. Fix a prime p and a positive integer ℓ . We consider two particular systems of linear forms. For $\mathbf{i} = (i_1, \dots, i_\ell) \in \mathbb{F}_p^\ell$, define

$$L_{\mathbf{i}}^\ell(x_1, \dots, x_\ell) := i_1 x_1 + \dots + i_\ell x_\ell.$$

Then define $\mathbf{L}^\ell := (L_{\mathbf{i}}^\ell)_{\mathbf{i} \in \mathbb{F}_p^\ell}$, the system of p^ℓ linear forms in ℓ variables that defines an ℓ -dimensional subspace.

Let $E_\ell \subset \mathbb{F}_p^\ell$ be the set of non-zero vectors whose first non-zero coordinate is 1. Then define $\overline{\mathbf{L}}^\ell := (L_{\mathbf{i}}^\ell)_{\mathbf{i} \in E_\ell}$, a system of $(p^\ell - 1)/(p - 1)$ linear forms in ℓ variables.

Note that unlike \mathbf{L}^ℓ , the system $\overline{\mathbf{L}}^\ell$ has finite complexity. For technical reasons, it will be convenient to reduce the removal lemma for general patterns to the case where all patterns are defined by a system of the form \mathbf{L}^ℓ . Then we reduce this to the following projective removal lemma where all patterns are defined by a system of the form $\overline{\mathbf{L}}^\ell$.

Theorem II.8 (Projective removal lemma). *Fix a prime p and a finite set \mathcal{S} equipped with an \mathbb{F}_p^\times -action. Let \mathcal{H} be a (possibly infinite) set consisting of \mathcal{S} -colored patterns over \mathbb{F}_p of the form $(\overline{\mathbf{L}}^\ell, \psi)$ where ℓ is some positive integer and $\psi: E_\ell \rightarrow \mathcal{S}$ is some map (see Definition II.7 for the definition of $\overline{\mathbf{L}}^\ell$ and E_ℓ). For every $\epsilon > 0$, there exists a finite subset $\mathcal{H}_\epsilon \subseteq \mathcal{H}$ and $\delta = \delta(\epsilon, \mathcal{H}) > 0$ such that the following holds. Let V be a finite-dimensional \mathbb{F}_p -vector space. If $f: V \rightarrow \mathcal{S}$ is a projective function with H -density at most δ for every $H \in \mathcal{H}_\epsilon$, then there exists a projective recoloring $g: V \rightarrow \mathcal{S}$ that agrees with f on all but an at most ϵ -fraction of V such that g has no generic H -instances for every $H \in \mathcal{H}$.*

III. PRELIMINARIES ON HIGHER-ORDER FOURIER ANALYSIS

A. Gowers norms and complexity

Definition III.1. Fix a prime p , a finite-dimensional \mathbb{F}_p -vector space V , and an abelian group G . Given a function $f: V \rightarrow G$ and a shift $h \in V$, define the **additive derivative** $D_h f: V \rightarrow G$ by

$$(D_h f)(x) := f(x + h) - f(x).$$

Given a function $f: V \rightarrow \mathbb{C}$ and a shift $h \in V$, define the **multiplicative derivative** $\Delta_h f: V \rightarrow \mathbb{C}$ by

$$(\Delta_h f)(x) := f(x + h) \overline{f(x)}.$$

Definition III.2. Fix a prime p and a finite-dimensional \mathbb{F}_p -vector space V . Given a function $f: V \rightarrow \mathbb{C}$ and $d \geq 1$, the **Gowers uniformity norm** $\|f\|_{U^d}$ is defined by

$$\|f\|_{U^d} := |\mathbb{E}_{x, h_1, \dots, h_d \in V} (\Delta_{h_1} \cdots \Delta_{h_d} f)(x)|^{1/2^d}.$$

Definition III.3. A system $\mathbf{L} = (L_1, \dots, L_m)$ of m linear forms in ℓ variables is **complexity at most d** if for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f_1, \dots, f_\ell: V \rightarrow [-1, 1]$ it holds that

$$|\Lambda_{\mathbf{L}}(f_1, \dots, f_\ell)| \leq \epsilon \quad \text{whenever} \quad \min_{1 \leq i \leq \ell} \|f_i\|_{U^{d+1}} \leq \delta. \quad (\text{III.1})$$

The **complexity** of \mathbf{L} is the smallest d such that the above holds, and infinite otherwise.

Remark III.4. The above definition is sometimes known as true complexity. It is known that a pattern (L_1, \dots, L_m) is complexity at most d if and only if $L_1^{d+1}, \dots, L_m^{d+1}$ are linearly independent as $(d+1)$ th order tensors [18], [24].

Let (L_1, \dots, L_m) be any pattern such that no form is identically zero and no two forms are linearly dependent. It is known (for example, because Cauchy-Schwarz complexity is an upper bound for true complexity [22]) that (L_1, \dots, L_m) has complexity at most $m - 2$.

It follows from the above discussion that the definition of complexity given in Definition III.3 agrees with the definition of finite complexity given in Definition II.6.

B. Non-classical polynomials and homogeneity

For ease of notation we write

$$\mathbb{U}_k := \frac{1}{p^k} \mathbb{Z}/\mathbb{Z} \subset \mathbb{R}/\mathbb{Z} \quad (\text{III.2})$$

through the paper.

Definition III.5. Fix a prime p , and a non-negative integer $d \geq 0$. Let V be a finite-dimensional \mathbb{F}_p -vector space. A **non-classical polynomial** of degree at most d is a map $P: V \rightarrow \mathbb{R}/\mathbb{Z}$ that satisfies

$$(D_{h_1} \cdots D_{h_{d+1}} P)(x) = 0$$

for all $h_1, \dots, h_{d+1}, x \in V$. The **degree** of P is the smallest $d > 0$ such that the above holds. The **depth** of P is the smallest $k \geq 0$ such that P takes values in a coset of \mathbb{U}_{k+1} .

In characteristic p , every non-classical polynomial of degree at most d has depth at most $\lfloor (d-1)/(p-1) \rfloor$ [33, Lemma 1.7(vi)].

Definition III.6. A **homogeneous non-classical polynomial** is a non-classical polynomial $P: V \rightarrow \mathbb{R}/\mathbb{Z}$ that also satisfies the following. For all $b \in \mathbb{F}_p$ there exists $\sigma_b^{(P)} \in \mathbb{Z}/p^{k+1}\mathbb{Z}$ such that $P(bx) = \sigma_b^{(P)} P(x)$ for all $x \in V$.

Lemma III.7 ([24, Lemma 3.3]). *Fix a prime p and integers $d > 0$ and $k \geq 0$ satisfying $k \leq \lfloor (d-1)/(p-1) \rfloor$. For each $b \in \mathbb{F}_p$ there exists $\sigma_b^{(d,k)} \in \mathbb{Z}/p^{k+1}\mathbb{Z}$ such that $\sigma_b^{(P)} = \sigma_b^{(d,k)}$ for all homogeneous non-classical polynomials P of degree d and depth k . Furthermore,*

for $b \neq 0$, the number $\sigma_b^{(d,k)}$ is uniquely determined by the following two properties:

- (i) $\sigma_b^{(d,k)} \equiv b^d \pmod{p}$
- (ii) $(\sigma_b^{(d,k)})^{p-1} = 1$ (in $\mathbb{Z}/p^{k+1}\mathbb{Z}$).

Theorem III.8 ([24, Theorem 3.4]). *Let P be a non-classical polynomial of degree d and depth k . Then P can be written as the sum of homogeneous non-classical polynomials of degree at most d and depth at most k .*

C. Polynomial factors

Definition III.9. Fix a prime p . Define

$$D_p := \{(d, k) \in \mathbb{Z}_{>0} \times \mathbb{Z}_{\geq 0} : k \leq \lfloor (d-1)/(p-1) \rfloor\},$$

and

$$\mathcal{I}_p := \left\{ I \in \mathbb{Z}_{\geq 0}^{D_p} : \sum_{(d,k) \in D_p} I_{d,k} < \infty \right\}.$$

We call $I \in \mathcal{I}_p$ a **parameter list**. For $I \in \mathcal{I}_p$, we write $\|I\| := p^{\sum_{(d,k) \in D_p} I_{d,k}}$ and $\deg I$ for the largest d such that $I_{d,k} \neq 0$ for some k . We add and subtract parameter list coordinatewise. For $I, I' \in \mathcal{I}_p$, we write $I \leq I'$ if $I_{d,k} \leq I'_{d,k}$ for all $(d, k) \in D_p$.

Definition III.10. For p a prime and $I \in \mathcal{I}_p$, define the **atom-indexing set** of I to be

$$A_I := \prod_{(d,k) \in D_p} \left(\frac{1}{p^{k+1}} \mathbb{Z}/\mathbb{Z} \right)^{I_{d,k}}. \quad (\text{III.3})$$

(Note that $|A_I| = \|I\|$.)

For $I, I' \in \mathcal{I}_p$ with $I \leq I'$, write $\pi: A_{I'} \rightarrow A_I$ for the standard projection map defined by

$$\pi \left(\left(a_{d,k}^i \right)_{\substack{(d,k) \in D_p \\ i \in [I'_{d,k}]} \right) \mapsto \left(a_{d,k}^i \right)_{\substack{(d,k) \in D_p \\ i \in [I_{d,k}]} \right). \quad (\text{III.4})$$

A_I is equipped with the following \mathbb{F}_p^\times -action:

$$c \cdot \left(a_{d,k}^i \right)_{\substack{(d,k) \in D_p \\ i \in [I_{d,k}]} := \left(\sigma_c^{(d,k)} a_{d,k}^i \right)_{\substack{(d,k) \in D_p \\ i \in [I_{d,k}]}}, \quad (\text{III.5})$$

where $\sigma_c^{(d,k)}$ is defined in Lemma III.7.

Definition III.11. Fix a prime p . Let V be a finite-dimensional \mathbb{F}_p -vector space and let $I \in \mathcal{I}_p$ be a parameter list. A **polynomial factor** on V with parameters I , denoted \mathfrak{B} , is a collection

$$\left(P_{d,k}^i \right)_{\substack{(d,k) \in D_p \\ i \in [I_{d,k}]}}$$

where $P_{d,k}^i$ is a homogeneous non-classical polynomial of degree d and depth k . We also use \mathfrak{B} to denote the map $\mathfrak{B}: V \rightarrow A_I$ defined by evaluation of the polynomials. We also associate to \mathfrak{B} the partition of V given

by the fibers of this map. The atoms of this partition are called the atoms of \mathfrak{B} . We write $\|\mathfrak{B}\| := \|I\|$ and $\deg \mathfrak{B} := \deg I$.

Note that if \mathfrak{B} is a polynomial factor on V with parameters I , then $\mathfrak{B}(cx) = c \cdot \mathfrak{B}(x)$ for all $c \in \mathbb{F}_p^\times$ and $x \in V$ where the \mathbb{F}_p^\times -action on A_I is defined in Eq. (III.5).

Definition III.12. Fix a prime p . Let V be a finite-dimensional \mathbb{F}_p -vector space and let $I, I' \in \mathcal{I}_p$ be two parameter lists. Let \mathfrak{B} and \mathfrak{B}' be two polynomial factors on V with parameters I and I' . We say that \mathfrak{B}' is a **refinement** of \mathfrak{B} if $I \leq I'$ and the lists of polynomials defining \mathfrak{B}' are extensions of the lists of polynomials defining \mathfrak{B} .

Note that if \mathfrak{B}' is a refinement of \mathfrak{B} , then $\mathfrak{B} = \pi \circ \mathfrak{B}'$ where $\pi: A_{I'} \rightarrow A_I$ is the projection defined in Eq. (III.4).

Definition III.13. Fix a prime p and integer $d \geq 0$. Let V be a finite-dimensional \mathbb{F}_p -vector space. For a non-classical polynomial $P: V \rightarrow \mathbb{R}/\mathbb{Z}$, define the **d -rank of P** , denoted $\text{rank}_d P$, to be the smallest integer r such that there exists non-classical polynomials $Q_1, \dots, Q_r: V \rightarrow \mathbb{R}/\mathbb{Z}$ of degree at most $d-1$ and a function $\Gamma: (\mathbb{R}/\mathbb{Z})^r \rightarrow \mathbb{R}/\mathbb{Z}$ such that $P(x) = \Gamma(Q_1(x), \dots, Q_r(x))$ for all $x \in V$.

For a polynomial factor \mathfrak{B} on V with parameters $I \in \mathcal{I}_p$, defined by a collection $(P_{d,k}^i)_{(d,k) \in D_p, i \in [I_{d,k}]}$ where $P_{d,k}^i$ is a homogeneous non-classical polynomial of degree d and depth k , we define the **rank of \mathfrak{B}** , denoted $\text{rank } \mathfrak{B}$, to be

$$\min_{\lambda \in \prod_{(d,k) \in D_p} (\mathbb{Z}/p^{k+1}\mathbb{Z})^{I_{d,k}}} \text{rank}_{d'} \left(\sum_{(d,k) \in D_p} \sum_{i=1}^{I_{d,k}} \lambda_{d,k}^i P_{d,k}^i \right)$$

where

$$d' := \min_{(d,k) \in D_p, i \in [I_{d,k}]} \deg(\lambda_{d,k}^i P_{d,k}^i).$$

D. Equidistribution and consistency sets

Definition III.14. Fix a prime p , integers $d > 0$ and $k \geq 0$ satisfying $k \leq \lfloor (d-1)/(p-1) \rfloor$, and a system $\mathbf{L} = (L_1, \dots, L_m)$ of m linear forms in ℓ variables. Define the **(d, k) -consistency set of \mathbf{L}** , denoted $\Phi_{d,k}(\mathbf{L})$, to be the subset of \mathbb{U}_{k+1}^m consisting of the tuples $\mathbf{a} = (a_1, \dots, a_m)$ such that there exists a finite-dimensional \mathbb{F}_p -vector space V , a homogeneous non-classical polynomial $P: V \rightarrow \mathbb{U}_{k+1}$ of degree d and depth k , and a tuple $\mathbf{x} \in V^\ell$ such that $a_i = P(L_i(\mathbf{x}))$ for all $i \in [m]$.

For a parameter list $I \in \mathcal{I}_p$, define the **I -consistency set of \mathbf{L}** to be the set of tuples $\mathbf{a} = (a_1, \dots, a_m) \in A_I^m$ such that for each $(d, k) \in D_p$ and $j \in [I_{d,k}]$ the tuple $((a_1)_{d,k}^j, \dots, (a_m)_{d,k}^j)$ lies in $\Phi_{d,k}(\mathbf{L})$.

Lemma III.15. Fix a prime p , integers $d > 0$ and $k \geq 0$ satisfying $k \leq \lfloor (d-1)/(p-1) \rfloor$, and a system $\mathbf{L} = (L_1, \dots, L_m)$ of m linear forms. The (d, k) -consistency set of \mathbf{L} is a subgroup of \mathbb{U}_{k+1}^m .

Theorem III.16 (Equidistribution [24, Theorem 3.10]). Fix a prime p , a positive integer $d > 0$, and a parameter $\epsilon > 0$. There exists $r_{\text{equi}}(p, d, \epsilon)$ such that the following holds. Let V be a finite-dimensional \mathbb{F}_p -vector space and let \mathfrak{B} be a polynomial factor on V with parameters I such that $\deg \mathfrak{B} \leq d$ and $\text{rank}(\mathfrak{B}) \geq r_{\text{equi}}(p, d, \epsilon)$. Then for a system of linear forms $\mathbf{L} = (L_1, \dots, L_m)$ consisting of m forms in ℓ variables, and a tuple of atoms $\mathbf{a} = (a_1, \dots, a_m) \in \Phi_I(\mathbf{L})$,

$$\left| \Pr_{\mathbf{x} \in V^\ell} (\mathfrak{B}(L_i(\mathbf{x})) = a_i \text{ for all } i \in [m]) - \frac{1}{|\Phi_I(\mathbf{L})|} \right| \leq \epsilon.$$

Remark III.17. To be completely correct, the statement given above follows by combining [24, Theorem 3.10] and [24, Corollary 2.13].

Note that the probability above is 0 if $\mathbf{a} \notin \Phi_I(\mathbf{L})$. We typically apply the above theorem with ϵ that decreases rapidly with $\|I\|$, for example, taking $\epsilon = 2\|I\|^{-m}$ and using the fact that $|\Phi_I(\mathbf{L})| \leq \|\mathfrak{B}\|^m$, we see that in this case the probability above is at least $1/(2|\Phi_I(\mathbf{L})|)$.

Consistency sets are often hard to compute exactly. The next lemma gives an exact relation on the sizes of consistency sets in a special case that occurs in this paper.

Definition III.18. Fix a prime p . A system \mathbf{L} of m linear forms in ℓ variables over \mathbb{F}_p is **full dimensional** if $|\Phi_{d,k}(\mathbf{L})| = |\Phi_{d,k}(\mathbf{L}^\ell)|$ for all $(d, k) \in D_p$ (recall the system \mathbf{L}^ℓ defined in Definition II.7 defines an ℓ -dimensional subspace).

Lemma III.19. Fix a prime p and a positive integer ℓ . Let $J \subseteq \mathbb{F}_p^\ell$ be a set that contains at least one vector in each direction (i.e., for each $\mathbf{i} \in \mathbb{F}_p^\ell$, there exists $\mathbf{j} \in J$ and $\mathbf{b} \in \mathbb{F}_p^\times$ such that $\mathbf{i} = \mathbf{b}\mathbf{j}$). Consider the system $\mathbf{L}_J := (L_{\mathbf{i}})_{\mathbf{i} \in J}$ of $|J|$ linear forms in ℓ variables (recall the linear form $L_{\mathbf{i}}$ defined in Definition II.7). Then \mathbf{L}_J is full dimensional.

As a special case of this result we see that the system $\overline{\mathbf{L}}^\ell$, defined in Definition II.7, is full dimensional.

E. Subatom selection functions

A situation that often occurs is the following. We have a polynomial factor \mathfrak{B} with parameters I and a refinement \mathfrak{B}' with parameters I' . We use the word atom to refer to the atoms of the partition induced by \mathfrak{B} ; these atoms are indexed by A_I . We use the word subatom to refer to the atoms of the partition induced by \mathfrak{B}' ; these atoms are indexed by $A_{I'}$. The projection map

$\pi: A_{I'} \rightarrow A_I$, defined in Eq. (III.4), maps a subatom to the atom that it is contained in.

We wish to designate one subatom inside each atom as special. This choice is given by a map $s: A_I \rightarrow A_{I'}$ that is a right inverse for π . In this paper we define a certain class of these maps that we call subatom selection functions that have several desirable properties.

First we define certain polynomials $P_{d,k}: \mathbb{F}_p \rightarrow \mathbb{U}_{k+1}$ for each $(d, k) \in D_p$. For each $i \in \{1, \dots, p-1\}$, there exists a homogeneous non-classical polynomial $\mathbb{F}_p \rightarrow \mathbb{U}_{k+1}$ of degree $k(p-1)+i$ and depth k in one variable. Let $P_{k(p-1)+i,k}$ be one such a polynomial. Finally, define

$$P_{(k+s)(p-1)+i,k+s} := P_{k(p-1)+i,k}$$

for all $i \in \{0, \dots, p-1\}$ and $s \geq 0$. This defines $P_{d,k}$ for each $(d, k) \in D_p$.

Definition III.20. Fix a prime p and parameter lists $I, I' \in \mathcal{I}_p$ satisfying $I \leq I'$. Let $c_{d,k}^{i,j} \in \mathbb{Z}/p^{k+1}\mathbb{Z}$ be arbitrary elements for $(d, k) \in D_p$ and $i \in [I_{1,0}]$ and $I_{d,k} < j \leq I'_{d,k}$. A **subatom selection function** is a map of the form $s_c: A_I \rightarrow A_{I'}$, defined by

$$[s_c(a)]_{d,k}^i = \begin{cases} a_{d,k}^i & \text{if } i \leq I_{d,k} \\ \sum_{j=1}^{I_{1,0}} c_{d,k}^{j,i} P_{d,k}(|a_{1,0}^j|) & \text{otherwise,} \end{cases}$$

where the maps $P_{d,k}: \mathbb{F}_p \rightarrow \mathbb{U}_{k+1}$ were defined in the preceding paragraph and $|\cdot|$ is the standard map $\mathbb{U}_1 \rightarrow \mathbb{F}_p$.

Lemma III.21. Fix a prime p , parameter lists $I, I' \in \mathcal{I}_p$ satisfying $I \leq I'$, and a subatom selection function $s_c: A_I \rightarrow A_{I'}$. The following hold:

- (i) $\pi \circ s_c = \text{Id}$ (where $\pi: A_{I'} \rightarrow A_I$ is defined in Eq. (III.4));
- (ii) for $a \in A_I$ and $b \in \mathbb{F}_p^\times$, we have $b \cdot s_c(a) = s_c(b \cdot a)$ (where the action of \mathbb{F}_p^\times on A_I and $A_{I'}$ is defined in Eq. (III.5));
- (iii) for every system \mathbf{L} of m linear forms and every consistent tuple of atoms $(a_1, \dots, a_m) \in \Phi_I(\mathbf{L})$, we have

$$(s_c(a_1), \dots, s_c(a_m)) \in \Phi_{I'}(\mathbf{L})$$

(see Definition III.14 for the definition of the consistency sets $\Phi_I(\mathbf{L})$ and $\Phi_{I'}(\mathbf{L})$).

IV. ARITHMETIC REGULARITY AND SUBATOM SELECTION

This section follows a fairly standard formula in the theory of regularity lemmas. We start with an inverse theorem, due to Tao and Ziegler [33]. Iterating the inverse theorem produces a weak regularity lemma, iterating the weak regularity lemma produces a regularity lemma, and iterating the regularity lemma gives a strong regularity lemma (Lemma IV.1). Finally we use the probabilistic

method applied to the output of the strong regularity lemma to produce the desired “subatom selection” result (Theorem IV.2).

The first two regularity lemmas we use are very similar to results in [8], [7], differing only in some technical details. The main innovation in this section is that Lemma IV.1 is much stronger than previous results. To accomplish this, we iterate our regularity lemma with the complexity parameter (i.e., degree of the non-classical polynomials) increasing at each step of the iteration. To our knowledge, this idea has not appeared previously in the literature.

This proceeding version only contains the statements of our strong arithmetic regularity lemma and the corresponding subatom selection result. The intermediate arithmetic regularity lemmas and their proofs can be found in the full version of the paper [34].

Notation and conventions: Recall that a polynomial factor \mathfrak{B} on a vector space V with parameters I gives rise to a partition (or σ -algebra) on V whose atoms are the fibers of the map $\mathfrak{B}: V \rightarrow A_I$. For a function $f: V \rightarrow \mathbb{C}$, we write $\mathbb{E}[f|\mathfrak{B}]: V \rightarrow \mathbb{C}$ for the projection of f onto the σ -algebra generated by \mathfrak{B} . Concretely, $\mathbb{E}[f|\mathfrak{B}](x)$ is defined to be the average of f over the atom of \mathfrak{B} which contains x .

Lemma IV.1 (Strong arithmetic regularity). *Fix a prime p , positive integers R, C_0, d_0 , a parameter $\zeta > 0$, non-increasing functions $\eta, \theta: \mathbb{Z}_{>0} \times \mathbb{Z}_{>0} \rightarrow (0, 1)$, and non-decreasing functions $d, r: \mathbb{Z}_{>0} \times \mathbb{Z}_{>0} \rightarrow \mathbb{Z}_{>0}$. There exist constants $C_{reg'''}(p, R, C_0, d_0, \zeta, \eta, \theta, d, r)$ and $D_{reg'''}(p, R, C_0, d_0, \zeta, \eta, \theta, d, r)$ and $r_{reg'''}(p, R, C_0, d_0, \zeta, \eta, \theta, d, r)$ such that the following holds. Let V be a finite-dimensional \mathbb{F}_p -vector space and let \mathfrak{B}_0 be a polynomial factor on V satisfying $\|\mathfrak{B}_0\| \leq C_0$ and $\deg \mathfrak{B}_0 \leq d_0$ and $\text{rank } \mathfrak{B}_0 \geq r_{reg'''}(p, R, C_0, d_0, \zeta, \eta, \theta, d, r)$. Given functions $f^{(1)}, \dots, f^{(R)}: V \rightarrow [0, 1]$, there exist a polynomial factor \mathfrak{B} and a refinement \mathfrak{B}' both on V with parameters I and I' with the following properties. There exists a decomposition*

$$f^{(\ell)} = f_{str}^{(\ell)} + f_{psr}^{(\ell)} + f_{sml}^{(\ell)}$$

for each $\ell \in [R]$ such that:

- (i) $f_{str}^{(\ell)} = \mathbb{E}[f^{(\ell)}|\mathfrak{B}']$ for each $\ell \in [R]$;
- (ii) $\|f_{psr}^{(\ell)}\|_{U^{d(\deg \mathfrak{B}, \|\mathfrak{B}\|)+1}} < \eta(\deg \mathfrak{B}', \|\mathfrak{B}'\|)$ for each $\ell \in [R]$;
- (iii) $f_{str}^{(\ell)}$ and $f_{str}^{(\ell)} + f_{sml}^{(\ell)}$ have range $[0, 1]$ and $f_{psr}^{(\ell)}$ and $f_{sml}^{(\ell)}$ have range $[-1, 1]$ for each $\ell \in [R]$;
- (iv) $\text{rank } \mathfrak{B} \geq r(\deg \mathfrak{B}, \|\mathfrak{B}\|)$ and $\text{rank } \mathfrak{B}' \geq r(\deg \mathfrak{B}', \|\mathfrak{B}'\|)$;
- (v) $\|f_{sml}^{(\ell)}\|_2 < \theta(\deg \mathfrak{B}, \|\mathfrak{B}\|)$ for each $\ell \in [R]$;

- (vi) for all but at most a ζ -fraction of $a \in A_{I'}$ it holds that

$$\left| \mathbb{E}_{x \in \mathfrak{B}'^{-1}(a)}[f^{(\ell)}(x)] - \mathbb{E}_{x \in \mathfrak{B}'^{-1}(\pi(a))}[f^{(\ell)}(x)] \right| < \zeta$$

for each $\ell \in [R]$ (recall the definition of the atom indexing sets from Eq. (III.3) and the projection map $\pi: A_{I'} \rightarrow A_I$ from Eq. (III.4));

- (vii) $\|\mathfrak{B}'\| \leq C_{reg'''}(p, R, C_0, d_0, \zeta, \eta, \theta, d, r)$ and $\deg \mathfrak{B}' \leq D_{reg'''}(p, R, C_0, d_0, \zeta, \eta, \theta, d, r)$.

Theorem IV.2 (Subatom selection). *Fix a prime p , positive integers R, c_0 , a parameter $\zeta > 0$, non-increasing functions $\eta, \theta: \mathbb{Z}_{>0} \times \mathbb{Z}_{>0} \rightarrow (0, 1)$, and non-decreasing functions $d, r: \mathbb{Z}_{>0} \times \mathbb{Z}_{>0} \rightarrow \mathbb{Z}_{>0}$. There exist constants $C_{reg}(p, R, c_0, \zeta, \eta, \theta, d, r)$ and $D_{reg}(p, R, c_0, \zeta, \eta, \theta, d, r)$ and $n_{reg}(p, c_0, \zeta)$ such that the following holds. Let V be a finite-dimensional \mathbb{F}_p -vector space satisfying $\dim V \geq n_{reg}(p, c_0, \zeta)$. Given functions $f^{(1)}, \dots, f^{(R)}: V \rightarrow [0, 1]$, there exist a polynomial factor \mathfrak{B} and a refinement \mathfrak{B}' both on V with parameters I and I' with the following properties. There exists a subatom selection function $s: A_I \rightarrow A_{I'}$ and a decomposition*

$$f^{(\ell)} = f_{str}^{(\ell)} + f_{psr}^{(\ell)} + f_{sml}^{(\ell)}$$

for each $\ell \in [R]$ such that:

- (i) $f_{str}^{(\ell)} = \mathbb{E}[f^{(\ell)}|\mathfrak{B}']$ for each $\ell \in [R]$;
- (ii) $\|f_{psr}^{(\ell)}\|_{U^{d(\deg \mathfrak{B}, \|\mathfrak{B}\|)+1}} < \eta(\deg \mathfrak{B}', \|\mathfrak{B}'\|)$ for each $\ell \in [R]$;
- (iii) $f_{str}^{(\ell)}$ and $f_{str}^{(\ell)} + f_{sml}^{(\ell)}$ have range $[0, 1]$ and $f_{psr}^{(\ell)}$ and $f_{sml}^{(\ell)}$ have range $[-1, 1]$ for each $\ell \in [R]$;
- (iv) $\text{rank } \mathfrak{B} \geq r(\deg \mathfrak{B}, \|\mathfrak{B}\|)$ and $\text{rank } \mathfrak{B}' \geq r(\deg \mathfrak{B}', \|\mathfrak{B}'\|)$;
- (v) for each $a \in A_I$ with $a_{1,0} \neq 0$, it holds that

$$\|f_{sml}^{(\ell)} \mathbf{1}_{\mathfrak{B}'^{-1}(s(a))}\|_2 < \theta(\deg \mathfrak{B}, \|\mathfrak{B}\|) \|\mathbf{1}_{\mathfrak{B}'^{-1}(s(a))}\|_2$$

for each $\ell \in [R]$;

- (vi) for all but at most a ζ -fraction of $a \in A_I$ it holds that

$$\left| \mathbb{E}_{x \in \mathfrak{B}'^{-1}(a)}[f^{(\ell)}(x)] - \mathbb{E}_{x \in \mathfrak{B}'^{-1}(s(a))}[f^{(\ell)}(x)] \right| < \zeta$$

for each $\ell \in [R]$;

- (vii) $I_{1,0} \geq c_0$;
- (viii) $\|\mathfrak{B}'\| \leq C_{reg}(p, R, c_0, \zeta, \eta, \theta, d, r)$ and $\deg \mathfrak{B}' \leq D_{reg}(p, R, c_0, \zeta, \eta, \theta, d, r)$.

V. PATCHING

To motivate the kind of results proved in this section consider the following result, which follows from an application of Ramsey’s theorem.

Let \mathcal{H} be a finite set of red/blue edge-colored graphs. There exists an integer $n_0 = n_0(\mathcal{H})$ such that the following holds. Either:

- (a) either the all-red coloring of K_n or the all-blue coloring of K_n contains no subgraph from \mathcal{H} for every n ; or
- (b) every 2-edge-coloring of K_n with $n \geq n_0$ contains a subgraph from \mathcal{H} .

We call such a statement a dichotomy result. The first main result of this section, Theorem V.6, is a dichotomy result for our setting. In our setting we consider colored labeled patterns instead of edge-colored subgraphs and instead of monochromatic colorings we have to consider so-called canonical colorings, defined below.

The second main result of this section, Theorem V.7, is our patching result, which is a supersaturation version of the dichotomy result.

Definition V.1. For a prime p , a finite set \mathcal{S} , and a parameter list $I \in \mathcal{I}_p$, an \mathcal{S} -colored I -labeled pattern over \mathbb{F}_p consisting of m linear forms in ℓ variables is a triple (\mathbf{L}, ψ, ϕ) given by:

- a system $\mathbf{L} = (L_1, \dots, L_m)$ of m linear forms in ℓ variables,
- a coloring $\psi: [m] \rightarrow \mathcal{S}$, and
- a labeling $\phi: [m] \rightarrow A_I$ (recall the definition of the atom-indexing set A_I from Eq. (III.3)).

Given a finite-dimensional \mathbb{F}_p -vector space V , a function $f: V \rightarrow \mathcal{S}$, and a polynomial factor \mathfrak{B} on V with parameters I , an (\mathbf{L}, ψ, ϕ) -instance in (f, \mathfrak{B}) is some $\mathbf{x} \in V^\ell$ such that $f(L_i(\mathbf{x})) = \psi(i)$ for all $i \in [m]$ and $\mathfrak{B}(L_i(\mathbf{x})) = \phi(i)$ for all $i \in [m]$. An instance is called **generic** if x_1, \dots, x_ℓ are linearly independent.

Definition V.2. For an \mathcal{S} -colored I -labeled pattern (\mathbf{L}, ψ, ϕ) consisting of m linear forms, a finite dimensional \mathbb{F}_p -vector space V , a function $f: V \rightarrow \mathcal{S}$, and a polynomial factor \mathfrak{B} on V with parameters I , define the (\mathbf{L}, ψ, ϕ) -density in (f, \mathfrak{B}) to be

$$\Lambda_{\mathbf{L}}(1_{f^{-1}(\psi(1)) \cap \mathfrak{B}^{-1}(\phi(1))}, \dots, 1_{f^{-1}(\psi(m)) \cap \mathfrak{B}^{-1}(\phi(m))}).$$

Given a set $X \subseteq V$, define the **relative density** of (\mathbf{L}, ψ, ϕ) in X to be

$$\frac{\Lambda_{\mathbf{L}}(f_1, \dots, f_m)}{\Lambda_{\mathbf{L}}(1_X, \dots, 1_X)}$$

where $f_i := 1_{X \cap f^{-1}(\psi(i)) \cap \mathfrak{B}^{-1}(\phi(i))}$.

Definition V.3. Define the **first non-zero coordinate function** $\text{fnz}: \mathbb{F}_p^n \rightarrow \mathbb{F}_p$ by $\text{fnz}(0, \dots, 0) := 0$ and $\text{fnz}(x_1, \dots, x_n) := x_k$ where $x_1 = \dots = x_{k-1} = 0$ and $x_k \neq 0$. Given a finite-dimensional \mathbb{F}_p -vector space V equipped with an isomorphism $\iota: V \xrightarrow{\sim} \mathbb{F}_p^n$, define the function $\text{fnz}_\iota: V \rightarrow \mathbb{F}_p$ by $\text{fnz}_\iota(x) := \text{fnz}(\iota(x))$.

Definition V.4. Fix a prime p , a finite set \mathcal{S} , a parameter list $I \in \mathcal{I}_p$, and a function $\xi: \mathbb{F}_p \times A_I \rightarrow \mathcal{S}$. For a finite-dimensional \mathbb{F}_p -vector space V equipped with an isomorphism $\iota: V \xrightarrow{\sim} \mathbb{F}_p^n$ and a polynomial factor \mathfrak{B} on

V with parameters I , define the ξ -canonical coloring $\Xi_{\xi, \iota, \mathfrak{B}}: V \rightarrow \mathcal{S}$ by $\Xi_{\xi, \iota, \mathfrak{B}}(x) := \xi(\text{fnz}_\iota(x), \mathfrak{B}(x))$. Furthermore, if \mathcal{S} is equipped with an \mathbb{F}_p^\times -action, say that ξ is **projective** if the same is true for every function $\Xi_{\xi, \iota, \mathfrak{B}}$. (Note that this property is equivalent to the condition that ξ preserves the action of \mathbb{F}_p^\times , i.e., $\xi(cx, c \cdot a) = c \cdot \xi(x, a)$ for all $c \in \mathbb{F}_p^\times$, all $x \in \mathbb{F}_p^n$, and all $a \in A_I$. Recall the action of \mathbb{F}_p^\times on A_I defined in Eq. (III.5).)

Definition V.5. Given a prime p , a finite set \mathcal{S} , a parameter list $I \in \mathcal{I}_p$, a function $\xi: \mathbb{F}_p \times A_I \rightarrow \mathcal{S}$, and a \mathcal{S} -colored I -labeled pattern $H = (\mathbf{L}, \psi, \phi)$, say that ξ **canonically induces** H if the following holds. There exists some $n \geq 0$ and a polynomial factor \mathfrak{B} on \mathbb{F}_p^n with parameters I such that there exists a generic H -instance in $(\Xi_{\xi, \text{Id}, \mathfrak{B}}, \mathfrak{B})$. For a finite set of \mathcal{S} -colored I -labeled patterns \mathcal{H} , say that ξ **canonically induces** \mathcal{H} if ξ canonically induces some $H \in \mathcal{H}$.

It is not hard to show that if ξ canonically induces H , then there exists a generic H -instance in $(\Xi_{\xi, \iota, \mathfrak{B}}, \mathfrak{B})$ for every V, ι, \mathfrak{B} as long as $\dim V$ and $\text{rank } \mathfrak{B}$ are large enough. Our first result is a strengthening of this: if every ξ canonically induces H , then there exists a generic H -instance in (f, \mathfrak{B}) for every $f: V \rightarrow \mathcal{S}$ and every \mathfrak{B} as long as $\dim V$ and $\text{rank } \mathfrak{B}$ are large enough.

Theorem V.6 (Dichotomy). Fix a prime p , a finite set \mathcal{S} with an \mathbb{F}_p^\times -action, a parameter list $I \in \mathcal{I}_p$, and a positive integer ℓ_0 . There exist constants $n_{\text{dich}} = n_{\text{dich}}(p, |\mathcal{S}|, I, \ell_0)$ and $r_{\text{dich}} = r_{\text{dich}}(p, |\mathcal{S}|, I, \ell_0)$ such that the following holds. Let \mathcal{H} be a finite set of \mathcal{S} -colored, I -labeled patterns each defined by a system of linear forms in at most ℓ_0 variables. Either:

- (a) there exists a projective $\xi: \mathbb{F}_p \times A_I \rightarrow \mathcal{S}$ that does not canonically induce \mathcal{H} ; or
- (b) for every finite-dimensional \mathbb{F}_p -vector space V satisfying $\dim V \geq n_{\text{dich}}$, every projective function $f: V \rightarrow \mathcal{S}$, and every polynomial factor \mathfrak{B} on V with parameters I which has rank at least r_{dich} , there is a generic H -instance in (f, \mathfrak{B}) for some $H \in \mathcal{H}$.

Theorem V.7 (Patching). Fix a prime p , a finite set \mathcal{S} with an \mathbb{F}_p^\times -action, parameter lists $I, I' \in \mathcal{I}_p$ satisfying $I \leq I'$, and a positive integer ℓ_0 . There exist constants $n_{\text{patch}} = n_{\text{patch}}(p, |\mathcal{S}|, I', \ell_0)$ and $\beta_{\text{patch}} = \beta_{\text{patch}}(p, |\mathcal{S}|, I, \ell_0) > 0$ and a non-decreasing function $r_{\text{patch}} = r_{\text{patch}}(p, |\mathcal{S}|, I, \ell_0): \mathbb{Z}_{>0} \times \mathbb{Z}_{>0} \rightarrow \mathbb{Z}_{>0}$ such that the following holds. Let \mathcal{H} be a finite set of \mathcal{S} -colored, I -labeled patterns such that each pattern is defined by a full dimension system of linear forms in at most ℓ_0 variables (recall Definition III.18). Either:

- (a) there exists a projective $\xi: \mathbb{F}_p \times A_I \rightarrow \mathcal{S}$ that does not canonically induce \mathcal{H} ; or

(b) for every finite-dimensional \mathbb{F}_p -vector space V satisfying $\dim V \geq n_{patch}$, every projective function $f: V \rightarrow \mathcal{S}$, every polynomial factor \mathfrak{B} on V with parameters I that satisfies $\text{rank } \mathfrak{B} \geq r_{patch}(\deg \mathfrak{B}, \|\mathfrak{B}\|)$, and every polynomial factor \mathfrak{B}' on V with parameters I' that refines \mathfrak{B} and satisfies $\text{rank } \mathfrak{B}' \geq r_{patch}(\deg \mathfrak{B}', \|\mathfrak{B}'\|)$, there is a pattern $H \in \mathcal{H}$ such that in (f, \mathfrak{B}) , the relative density of H in $\mathfrak{B}'^{-1}(A_I \times \{0\})$ is at least β_{patch} .

VI. CONCLUSION

The projective removal lemma, Theorem II.8, can be deduced from the subatom selection and patching results, Theorem IV.2 and Theorem V.7, together with a compactness argument originally due to Alon and Shapira. The main removal lemma, Theorem II.4, can be reduced to the projective removal lemma. Finally the main property testing results, Theorem I.3 and Theorem I.4, follow directly from the main removal lemma. All of these proofs can be found in the full version of the paper [34].

REFERENCES

- [1] Noga Alon, Richard A. Duke, Hanno Lefmann, Vojtěch Rödl, and Raphael Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms **16** (1994), 80–109.
- [2] Noga Alon, Eldar Fischer, Michael Krivelevich, and Mario Szegedy, *Efficient testing of large graphs*, Combinatorica **20** (2000), 451–476.
- [3] Noga Alon, Tali Kaufman, Michael Krivelevich, Simon Litsyn, and Dana Ron, *Testing Reed-Muller codes*, IEEE Trans. Inform. Theory **51** (2005), 4032–4039.
- [4] Noga Alon and Asaf Shapira, *A characterization of the (natural) graph properties testable with one-sided error*, SIAM J. Comput. **37** (2008), 1703–1727.
- [5] Vitaly Bergelson, Terence Tao, and Tamar Ziegler, *An inverse theorem for the uniformity seminorms associated with the action of \mathbb{F}_p^∞* , Geom. Funct. Anal. **19** (2010), 1539–1596.
- [6] Arnab Bhattacharyya, Victor Chen, Madhu Sudan, and Ning Xie, *Testing linear-invariant non-linear properties*, Theory Comput. **7** (2011), 75–99.
- [7] Arnab Bhattacharyya, Eldar Fischer, Hamed Hatami, Pooya Hatami, and Shachar Lovett, *Every locally characterized affine-invariant property is testable*, STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing, ACM, New York, 2013, pp. 429–435.
- [8] Arnab Bhattacharyya, Eldar Fischer, and Shachar Lovett, *Testing low complexity affine-invariant properties*, Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, PA, 2012, pp. 1337–1355.
- [9] Arnab Bhattacharyya, Elena Grigorescu, and Asaf Shapira, *A unified framework for testing linear-invariant properties*, Random Structures Algorithms **46** (2015), 232–260.
- [10] Manuel Blum, Michael Luby, and Ronitt Rubinfeld, *Self-testing/correcting with applications to numerical problems*, Proceedings of the 22nd Annual ACM Symposium on Theory of Computing (Baltimore, MD, 1990), vol. 47, 1993, pp. 549–595.
- [11] David Conlon and Jacob Fox, *Graph removal lemmas*, Surveys in combinatorics 2013, London Math. Soc. Lecture Note Ser., vol. 409, Cambridge Univ. Press, Cambridge, 2013, pp. 1–49.
- [12] Jacob Fox, Jonathan Tidor, and Yufei Zhao, *Induced arithmetic removal: complexity 1 patterns over finite fields*, arXiv:1911.03427.
- [13] Zoltán Füredi, *Extremal hypergraphs and combinatorial geometry*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), Birkhäuser, Basel, 1995, pp. 1343–1352.
- [14] Oded Goldreich, Shafi Goldwasser, and Dana Ron, *Property testing and its connection to learning and approximation*, J. ACM **45** (1998), 653–750.
- [15] Oded Goldreich and Dana Ron, *On proximity-oblivious testing*, SIAM J. Comput. **40** (2011), 534–566.
- [16] W. Timothy Gowers, *A new proof of Szemerédi’s theorem*, Geom. Funct. Anal. **11** (2001), 465–588.
- [17] W. Timothy Gowers, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Ann. of Math. (2) **166** (2007), 897–946.
- [18] W. Timothy Gowers and Julia Wolf, *Linear forms and higher-degree uniformity for functions on \mathbb{F}_p^n* , Geom. Funct. Anal. **21** (2011), 36–69.
- [19] Ben Green, *A Szemerédi-type regularity lemma in abelian groups, with applications*, Geom. Funct. Anal. **15** (2005), 340–376.
- [20] Ben Green and Tom Sanders, *Fourier uniformity on subspaces*, arXiv:1510.08739.
- [21] Ben Green and Terence Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. (2) **167** (2008), 481–547.
- [22] Ben Green and Terence Tao, *Linear equations in primes*, Ann. of Math. (2) **171** (2010), 1753–1850.
- [23] Ben Green, Terence Tao, and Tamar Ziegler, *An inverse theorem for the Gowers $U^{s+1}[N]$ -norm*, Ann. of Math. (2) **176** (2012), 1231–1372.
- [24] Hamed Hatami, Pooya Hatami, and Shachar Lovett, *General systems of linear forms: equidistribution and true complexity*, Adv. Math. **292** (2016), 446–477.
- [25] Hamed Hatami, Pooya Hatami, and Shachar Lovett, *Higher-order fourier analysis and applications*, Foundations and Trends® in Theoretical Computer Science **13** (2019), 247–448.
- [26] Tali Kaufman and Madhu Sudan, *Algebraic property testing: the role of invariance*, STOC'08, ACM, New York, 2008, pp. 403–412.
- [27] Daniel Král', Oriol Serra, and Lluís Vena, *A removal lemma for systems of linear equations over finite fields*, Israel J. Math. **187** (2012), 193–207.
- [28] Vojtěch Rödl, Brendan Nagle, Jozef Skokan, Mathias Schacht, and Yoshiharu Kohayakawa, *The hypergraph regularity method and its applications*, Proc. Natl. Acad. Sci. USA **102** (2005), 8109–8113.
- [29] Vojtěch Rödl and Mathias Schacht, *Property testing in hypergraphs and the removal lemma [extended abstract]*, STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing, ACM, New York, 2007, pp. 488–495.
- [30] Imre Z. Ruzsa and Endre Szemerédi, *Triple systems with no six points carrying three triangles*, Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II, Colloq. Math. Soc. János Bolyai, vol. 18, North-Holland, Amsterdam-New York, 1978, pp. 939–945.
- [31] Asaf Shapira, *A proof of Green’s conjecture regarding the removal properties of sets of linear equations*, J. Lond. Math. Soc. (2) **81** (2010), 355–373.
- [32] Terence Tao and Tamar Ziegler, *The inverse conjecture for the Gowers norm over finite fields via the correspondence principle*, Anal. PDE **3** (2010), 1–20.
- [33] Terence Tao and Tamar Ziegler, *The inverse conjecture for the Gowers norm over finite fields in low characteristic*, Ann. Comb. **16** (2012), 121–188.
- [34] Jonathan Tidor and Yufei Zhao, *Testing linear-invariant properties [full version]*, arXiv:1911.06793.