THE INVERSE SIEVE PROBLEM IN HIGH DIMENSIONS

MIGUEL N. WALSH

Abstract. We show that if a big set of integer points $S \subseteq [0,N]^d$, $d > 1$, occupies few residue classes mod p for many primes p, then it must essentially lie in the solution set of some polynomial equation of low degree. This answers a question of Helfgott and Venkatesh.

1. Introduction

One of the main topics of study in analytic number theory is the distribution of sets of integers in residue classes. Examples abound, but folkloric ones include Dirichlet's theorem, which tells us that the primes are uniformly distributed along primitive residue classes, and the open problem of determining how large may the least quadratic non-residue be.

On the other hand, in the expanding subject of arithmetic combinatorics, much of the focus has been in establishing what is known as *inverse theorems* in which one starts with a set having a specific arithmetic property and wishes to use this information to give a characterization of the set. Notable examples of this include Freiman type theorems (see for instance [6, 10, 17] and the survey [4]), inverse theorems for the Gowers norm [7] and the inverse Littlewood-Offord theory [18, 19].

This paper is concerned with the problem of connecting both lines of inquiry by establishing an inverse theorem for the distribution of sets in residue classes. Since we would expect a random set to be fairly well distributed, the main question here is whether a set occupying very few residue classes for many primes p has to have some specific structure. The remarkable observation that this might indeed be the case is due to Croot and Elsholtz [3] and Helfgott and Venkatesh [11]. Writing $[N]$ for the set of integers $\{0,\dots,N\}$ their observation can be resumed in the following principle:

Inverse Sieve Problem. *Suppose a set $S \subseteq [N]^d$ occupies very few residue classes mod p for many primes p. Then, either S is small, or it possesses some strong algebraic structure.*

There is a good reason why such inverse sieve results are of much interest in number theory. One of the main features of sieve theory is the uniformity of its results, which is a consequence of the fact that sieves only take into account the cardinality of the classes occupied by the set. However, a clear drawback of this is that the bounds thus obtained are limited to what happens in extremal cases. By stating that such extremal sets must have a very specific structure, inverse results should allow one to retain the uniformity of the sieve while providing much stronger bounds. The reader may consult the book of Kowalski [13, §2.5] for further discussion of the potential applications of this phenomenon and [4] for applications of similar classifications in arithmetic combinatorics.

The author was partially supported by a CONICET doctoral fellowship.

In this paper we give a satisfactory answer to the inverse sieve problem for every $d \geq 2$. In order to discuss our results suppose we are given a big integer set $S \subseteq [N]^d$ occupying $O(p^{d-1})$ residue classes in $(\mathbb{Z}/p\mathbb{Z})^d$ for many primes p. What does this imply about S? By the Lang-Weil inequality, we know that this condition is satisfied by the set of integer points of a proper algebraic variety of small degree and one would expect a partial converse to also hold. That is, that any big set $S \subseteq [N]^d$ occupying only that many residue classes for every prime p should essentially be contained inside the solution set of a polynomial of low degree. When $d = 1$ this follows from Gallagher's larger sieve [8] (not to be confused with the conjecture discussed in §6.3). The case $d = 2$ was proven by Helfgott and Venkatesh in [11], by applying the Bombieri-Pila determinant method [1] to obtain a two-dimensional generalization of the larger sieve. Although their methods are only capable of handling the case $d \leq 2$, they conjectured that such an inverse theorem should in fact hold for every dimension d. In this paper we introduce a different approach and use it to answer this question by giving the following best possible result.

Theorem 1.1. *Let $0 \leq k < d$ be integers and let $\varepsilon, \alpha, \eta > 0$ be positive real numbers. Then, there exists a constant C depending only on the above parameters, such that for any set $S \subseteq [N]^d$ occupying less than αp^k residue classes for every prime p at least one of the following holds:*

(i) (S is small) $|S| \ll_{d,k,\varepsilon,\alpha} N^{k-1+\varepsilon}$,

(ii) (S is strongly algebraic) There exists a polynomial $f \in \mathbb{Z}[x_1, \ldots, x_d]$ of degree at most C and coefficients bounded by N^C vanishing at more than $(1-\eta)|S|$ points of S.

Theorem 1.1 is sharp. Indeed, the reader may consult Section §5 for examples of sets of size $|S| \gg N^{k-1}$ occupying less than p^k residue classes for every prime p but possessing no algebraic structure. On the other hand, we only need to require from S that it occupies few residue classes for sufficiently many small primes (see Theorem 2.4). More generally, we will show in Theorem 6.1 that assuming some necessary regularity conditions, every set of size $\gg N^{\varepsilon}$ occupying few residue classes for many primes p must satisfy condition (ii). In Section §6.2 we shall give an easy application of this generalization to the characterization of functions preserving some structure when reduced to prime moduli.

Taking $d = 2$ in Theorem 1.1 we recover the result of [11]. Actually, the methods of Helfgott and Venkatesh are capable of handling the case $k = 1$ of Theorem 1.1, that is, when S is assumed to occupy only $O(p)$ residue classes. However, the approach fails as soon as the set occupies more than $p \log p$ classes. The reason for this is that their method, as well as the larger sieve itself, is in essence a counting argument (see §3.1) and therefore needs the *number* of classes occupied by S to be small, while the high dimensional setting requires us to take advantage of the local *density* of S being small. This type of obstacle is not specific to the problem at hand, but arises whenever one tries to extend this kind of sieves to higher dimensional settings (see [12, Remark 3] for some discussion). So while we do make use of the larger sieve, in order to establish Theorem 1.1 we need to introduce an approach that overcomes this difficulty by taking advantage of the structure of the set and which we believe to be applicable in more general situations.

The rest of the paper is organized as follows. After setting up some notation, in Section §2 we state and discuss Proposition 2.2, which is the main ingredient of the paper, and use it to deduce Theorem 1.1. Roughly speaking, this proposition says that every set satisfying hypothesis similar to those of Theorem 1.1 admits a subset of size $O(r^k)$ such that if a polynomial identity of degree r holds at this set

then it must also hold at a positive proportion of the points of S. Then, in Section §3, we review some facts about the larger sieve and apply them to obtain a key uniformization lemma. Using this, the proof of Proposition 2.2 is carried out in Section §4. Finally, in §5 we construct several examples showing that our results are sharp, while in §6 we discuss further consequences of our methods as well as the remaining case ($d = 1$) of the inverse sieve problem.

Acknowledgment. The author would like to thank his advisor Román Sasyk for several comments and suggestions during the preparation of the paper. He would also like to thank the two anonymous referees for their helpful suggestions.

2. A conditional proof of Theorem 1.1

2.1. **General notation.** We now fix some notation. By $O_{c_1,\dots,c_k}(X)$ we shall mean a quantity which is bounded by $C_{c_1,\dots,c_k}X$ where $C_{c_1,\dots,c_k}$ is some finite positive constant depending on $c_1,\dots,c_k$. Also, we shall write $Y \ll_{c_1,\dots,c_k} X$ to mean $|Y| = O_{c_1,\dots,c_k}(X)$. However, since we will generally be concerned with the study of a set S satisfying the hypothesis of Proposition 2.2 for some parameters d, h, κ and ε as in the statement of that proposition, we will free up some notation by assuming that all implied constants in the $O, \ll$ notation *always* depend on these parameters even though this may not be explicitly stated. So for instance $Y \ll_\eta X$ stands for $Y \ll_{\eta,d,h,\kappa,\varepsilon} X$. Throughout the paper we will let the letter c denote a small positive constant whose exact value may vary at each occurrence.

Given a statement $\phi(x)$ with respect to an element $x \in [N]^d$ we will write $\mathbf{1}_{\phi(x)}$ for the function which equals 1 if $\phi(x)$ is true and 0 otherwise. Also, we shall write $\pi_i : \mathbb{Z}^d \to \mathbb{Z}$, $1 \le i \le d$, for the projection to the ith coordinate.

The letter p will always refer to a prime number. We write $\mathcal{P}$ for the set of primes and given any magnitude Q, we denote $\mathcal{P}(Q)$ the set of primes $p \le Q$. Since we will usually need to consider the weight $\frac{\log p}{p}$ over $\mathcal{P}$, for a finite subset $P \subseteq \mathcal{P}$ we write $w(P) := \sum_{p\in P} \frac{\log p}{p}$. We shall use the estimates $w(\mathcal{P}(Q)) = \log Q + O(1)$ and $\sum_{p\in\mathcal{P}(Q)} \log p \sim Q$ without explicit mention.

2.2. **Characteristic sets.** The purpose of this section is to state Proposition 2.2 which is the key ingredient of the paper and use it to derive Theorem 1.1. What this Proposition essentially says, is that for any ill-distributed set S as in the statement of Theorem 1.1, one may find a very small "characteristic" subset $A \subseteq S$ such that if a small polynomial vanishes at A then it also vanishes at a positive proportion of S. The task of proving Theorem 1.1 is thus reduced to that of finding a polynomial which vanishes at A, and this will always be possible since A is small.

Before proceeding, we need to define exactly what we mean for a polynomial to be small. Given a parameter N and some integer $d > 0$ by an r-polynomial, for a positive integer r, we shall mean any polynomial f with integer coefficients satisfying $|f(n)| < N^{3r}$ for every $n \in [N]^d$. The exponent $3r$ is chosen in order to guarantee that if N is sufficiently large in terms of r and d, then a polynomial $f \in \mathbb{Z}[x_1,\dots,x_d]$ of degree at most r, with coefficients bounded in absolute value by N^r, is an r-polynomial. This leads us to the following definition.

Definition 2.1. Let $0 < \delta \le 1$ be a positive real number and $r > 0$ some integer. We say a subset A of a set S is (r,δ)-characteristic for S if we can find some subset $A \subseteq B \subseteq S$ of size $|B| \ge \delta|S|$ such that whenever an r-polynomial vanishes at A, then it also vanishes at B.

We can now state Proposition 2.2 which says that ill-distributed sets always admit small characteristic subsets.

Proposition 2.2. *Let $d, h \geq 1$ be arbitrary integers and $\varepsilon > 0$ some positive real number. Set $Q = N^{\frac{\varepsilon}{2d}}$ and let $P \subseteq \mathcal{P}(Q)$ satisfy $w(P) \geq \kappa \log Q$ for some $\kappa > 0$. Also, let r be an arbitrary positive integer. Suppose $S \subseteq [N]^d$ is a set of size $|S| \gg N^{d-h-1+\varepsilon}$ occupying at most αp^{d-h} residue classes mod p for every prime $p \in P$ and some $\alpha > 0$. Then, if N is sufficiently large, there exists a set $A \subseteq S$ of size $|A| = O(r^{d-h})$ which is (r, δ)-characteristic for S, for some $\delta > 0$ which depends on $d, h, \kappa, \varepsilon$ but is independent of S, N or r.*

Remarks. The exact value of Q in the above statement is irrelevant and may be replaced by any small power of N. The reason why we have made the change of variables $h := d - k$ with respect to Theorem 1.1 is that in the arguments to follow we shall always set the quantity h to be fixed and induct on d. We believe it is simpler to introduce this change of notation at an early stage.

To see why such a result might be expected consider some polynomial f vanishing at an integer point x. Since polynomials descend to congruence classes, this means that for any other integer y satisfying $y \equiv x (\mathrm{mod}\ p)$ for a prime p, we will have $p|f(y)$. Thus, if we are given a set S which occupies very few residue classes, one may then hope to find a small subset A such that given some $y \in S$ there are a lot of primes p for which $y \equiv x (\mathrm{mod}\ p)$ for some $x \in A$. It would then follow that if a polynomial vanishes at A then there would be many primes p dividing $f(y)$. If furthermore f is small, then this can only hold if $f(y) = 0$. Notice that this is similar to the general idea of the larger sieve, where one uses the fact that S occupies few residue classes mod p to conclude the existence of too many pairs of elements of S occupying the same class, and contrasts this with the fact that no fixed pair of distinct bounded integers can occupy the same residue class for many primes (see §3.1).

On the other hand, the size hypothesis on S is necessary. For instance, one may construct small (logarithmic size) sets $S \subseteq [N]$ as in [11, §4.3] which occupy few residue classes for large moduli just because they are small, but which however have at most one element in each residue class, making the above argument unviable in this situation. Furthermore, it is clear that a similar pathology occurs in higher dimensions, by considering for instance the product set $S \times [N]$. For the general construction of this type of sets and to see that in fact one cannot take $\varepsilon = 0$ in Proposition 2.2 the reader is referred to §5.

In order to deduce Theorem 1.1 from Proposition 2.2 we will need to find a polynomial which vanishes at a specific set of points. This will be accomplished in a standard way by means of Siegel's lemma.

Lemma 2.3 (Siegel)**.** *Suppose we are given a system of m linear equations*

$$\sum_{j=1}^{n} a_{ij}\beta_j = 0 \quad \forall 1 \leq i \leq m,$$

in n unknowns $(\beta_1, \ldots, \beta_n)$, $n > m$, where the coefficients (a_{ij}) are integers not all equal to 0 and bounded in magnitude by some constant C. Then, the above system has a non trivial integer solution $(\beta_1, \ldots, \beta_n)$ with $|\beta_j| \leq 1 + (Cn)^{m/(n-m)}$ for all $1 \leq j \leq n$.

Proof of Theorem 1.1 assuming Proposition 2.2. Let the hypothesis be as in the statement of Theorem 1.1 and write $h := d - k$. Assume condition (i) fails, so that $|S| \gg N^{d-h-1+\varepsilon}$. We claim that for any given integer r there exists a set $A \subseteq S$ of size $|A| = O_\eta(r^{d-h})$ which is $(r, 1-\eta)$-characteristic for S, provided N is sufficiently large. To see this we begin by noticing that Proposition 2.2 implies the existence of some $\delta \gg 1$ such that for every subset $S' \subseteq S$ with $|S'| \geq \eta|S|$

there exists a set $A' \subseteq S'$ of size $|A'| = O_\eta(r^{d-h})$ which is (r,δ)-characteristic for S'. From now on we fix this value of δ. Let A_0 be such a characteristic subset for S and let B_0 consist of those elements of S which vanish at every r-polynomial that vanishes at A_0, so in particular $|B_0| \geq \delta|S|$. If $\delta \geq 1-\eta$ we are done, otherwise we have that $S_1 := S \setminus B_0$ satisfies $|S_1| \geq \eta|S|$ and therefore contains a characteristic subset $A_1 \subseteq S_1$ as above. If we now let B_1 denote those points of S_1 vanishing at every r-polynomial that vanishes at A_1 we see that either we get the claim with $A = A_0 \cup A_1$ or the set $S_2 := S_1 \setminus B_1$ satisfies $\eta|S| \leq |S_2| \leq (1-\delta)^2|S|$. After iterating this process j times we see that if the set $A = \bigcup_{i=0}^{j-1} A_i$ is not $(r,1-\eta)$-characteristic for S then we can find some $S_j \subseteq S$ with $\eta|S| \leq S_j \leq (1-\delta)^j|S|$. Since this last possibility cannot hold for some large $j = O_\eta(1)$, the claim follows.

Now it only remains to find some r-polynomial f which vanishes at A and which is of the form given in Theorem 1.1. Notice that this is plausible since the size of A is $\ll_\eta r^{d-h}$ while an r-polynomial has $\sim r^d$ degrees of freedom. We now make this rigorous by means of Siegel's lemma. Thus, we may assume $d|r$ and consider the system of $|A|$ linear equations in $\left(\frac{r}{d}+1\right)^d$ unknowns given by

$$\sum_{\mathbf{i}=\{i_1,\ldots,i_d\}\leq r/d} \beta_{\mathbf{i}} a^{\mathbf{i}} = 0 \ \forall a \in A, \tag{2.1}$$

where $\mathbf{i} \leq l$ stands for $i_j \leq l$ for all $1 \leq j \leq d$ and where we use the multi-index notation $a^{\mathbf{i}} = a_1^{i_1} \ldots a_d^{i_d}$ for $a = (a_1, \ldots, a_d)$. Notice that $|a^{\mathbf{i}}| \leq N^r$ for every $\mathbf{i}$ and that a solution $(\beta_{\mathbf{i}})$ of (2.1) corresponds to the coefficients of a polynomial vanishing at A. If we now choose $r = O_\eta(1)$ large enough so that $\left(\frac{r}{d}+1\right)^d > 3|A|$ it follows by Siegel's lemma that there exists an integer solution $(\beta_{\mathbf{i}})$ to (2.1) with $|\beta_{\mathbf{i}}| \ll_r N^{r/2} \leq N^r$ provided N is sufficiently large. We thus see that the polynomial $f := \sum_{\mathbf{i}\leq r/d} \beta_{\mathbf{i}} x^{\mathbf{i}}$ is of the desired form (assuming again that N is sufficiently large) and, taking $C = r$, this concludes the proof of Theorem 1.1. □

Notice that we have actually proved the following slight strengthening of Theorem 1.1 in which the set S is only required to be badly distributed in a dense subset of the primes.

Theorem 2.4. *Let $0 \leq k < d$ be integers and let $\varepsilon, \eta > 0$ be positive real numbers. Set $Q = N^{\frac{\varepsilon}{2d}}$ and let $P \subseteq \mathcal{P}(Q)$ satisfy $w(P) \geq \kappa \log Q$ for some $\kappa > 0$. Suppose $S \subseteq [N]^d$ is a set of size $|S| \gg N^{k-1+\varepsilon}$ occupying at most αp^k residue classes mod p for every prime $p \in P$ and some $\alpha > 0$. Then there exists a polynomial $f \in \mathbb{Z}[x_1, \ldots, x_d]$ of degree $O_\eta(1)$ and coefficients bounded by $N^{O_\eta(1)}$ which vanishes at more than $(1-\eta)|S|$ points of S.*

Remark. Since we have already mentioned that the exact value of Q in Proposition 2.2 is irrelevant, it follows that Theorem 2.4 also holds with Q any small power of N.

3. Applying the larger sieve in high dimensions

3.1. **A review of the larger sieve.** We will now quickly review some facts about Gallagher's larger sieve and use them to prove two easy lemmas which we shall need later. For further discussion of the larger sieve and its consequences the reader may consult [2, Section 2.2] and of course Gallagher's original paper [8].

Before proceeding we need to state some further notation that will be used in this and the next sections. When studying a set $S \subseteq [N]^d$ we will denote by $[S]_p$ the set of residue classes mod p occupied by S. Given such a set S, we shall be largely concerned with how many elements of S belong to a given residue class, so it

is important for us to have a specific notation for this subset. Thus, given a residue class $\mathbf{a} = (a_1, \dots, a_d) (\mathrm{mod}\ p)$ we write $S(\mathbf{a}; p)$ to refer to those elements of S which are congruent to $\mathbf{a} (\mathrm{mod}\ p)$. Moreover, we shall sometimes consider some $a \in \mathbb{Z}/p\mathbb{Z}$ and write $S(a; p)$ for those elements of S having their first coordinate congruent to $a \pmod p$. Since we will always use the bold font $\mathbf{a}$ to denote a vector residue class and since where this class lives shall be clear from the context altogether, we believe the similarity of both notations will not cause any confusion. Finally, if p is fixed, we may simply write $S(\mathbf{a})$ and $S(a)$ for the above sets.

Fix now a set S and consider some parameter Q. The main idea of the larger sieve is to count in two different ways the number of distinct pairs $x, y \in S$ and primes $p \leq Q$ such that $x \equiv y (\mathrm{mod}\ p)$. Given two such integers $x, y \in [N]$ it is clear that those primes for which they are congruent are exactly those dividing $|x - y| \leq N$ and therefore

$$\sum_{p \leq Q} \sum_{\substack{x,y \in S \\ x \neq y}} \mathbf{1}_{x \equiv y (\mathrm{mod}\ p)} \log p \leq |S|^2 \log N. \tag{3.1}$$

On the other hand, we have that the left hand side of (3.1) equals

$$\sum_{p \leq Q} \sum_{a (\mathrm{mod}\ p)} |S(a; p)|^2 \log p - |S| \sum_{p \leq Q} \log p. \tag{3.2}$$

Notice that the above argument also works if $S \subseteq [N]^d$ since if p is a prime for which $x \equiv y (\mathrm{mod}\ p)$ then p must divide $|\pi_1(x) - \pi_1(y)|$ which is bounded by N.

As an example, we have the following result due to Gallagher [8]. Suppose we are given a set $S \subseteq [N]$ occupying at most αp residue classes, then the Cauchy-Schwarz inequality implies

$$\sum_{a (\mathrm{mod}\ p)} |S(a; p)|^2 \geq \frac{1}{\alpha p} |S|^2.$$

Combining this with (3.1) and (3.2) we obtain

$$\frac{1}{\alpha} \log Q + O\left(\frac{|Q|}{|S|}\right) \leq \log N + O(1).$$

Taking $Q = |S|$ we conclude that $|S| \ll_\alpha N^\alpha$.

For the purposes of this paper, we need to apply Gallagher's sieve in a slightly more general context. Precisely, we will use the following lemma which is also classical.

Lemma 3.1. *Let $X \subseteq [N]$ be some set of integers and set $Q = N^\gamma$ for some $\gamma > 0$. Let $c_1, c_2 > 0$ be positive real numbers. Suppose there is a set of primes $P \subseteq \mathcal{P}(Q)$ with $w(P) \geq c_1 \log Q$ such that for every $p \in P$ there are at least $c_2|X|$ elements of X lying in at most αp residue classes for some $\alpha > 0$ independent of p. Then, if α is sufficiently small in terms of c_1, c_2 and γ, it must be $|X| < Q$.*

Proof. Again, we count the number of pairs $x, y \in X$ and $p \in P$ with $x \equiv y (\mathrm{mod}\ p)$. On one hand, we have as before that

$$\sum_{p \in P} \sum_{\substack{x,y \in X \\ x \neq y}} \mathbf{1}_{x \equiv y (\mathrm{mod}\ p)} \log p \leq |X|^2 \log N. \tag{3.3}$$

On the other hand, using the Cauchy-Schwarz inequality we see that our hypothesis on X implies

$$\sum_{a (\mathrm{mod}\ p)} |X(a; p)|^2 \geq \frac{1}{\alpha p} (c_2 |X|)^2,$$

from where it follows that the left hand side of (3.3) is at least

$$\frac{c_1 c_2^2}{\alpha}|X|^2 \log Q + O(|Q||X|).$$

It is then clear that if α is sufficiently small, then the only way for (3.3) to hold is to have $|X| < Q$. □

Finally, we prove the following easy consequence of the larger sieve which already handles the case $d = h$ of Proposition 2.2.

Lemma 3.2. *Let $Q = N^\gamma$ for some $\gamma > 0$ and let $P \subseteq \mathcal{P}(Q)$ be some set of primes with $w(P) \geq c_1 \log Q$ for some $c_1 > 0$. Let $S \subseteq [N]^d$ occupy less than c_2 residue classes mod p for every prime $p \in P$ and some constant c_2. Then $|S| = O_{c_1,c_2,\gamma}(1)$.*

Proof. Gallagher's sieve implies in this case

$$\log N \geq \left(\frac{1}{c_2} - \frac{1}{|S|}\right) \sum_{p \in P} \log p \gg \left(\frac{1}{c_2} - \frac{1}{|S|}\right) N^{\gamma c_1},$$

and clearly, for sufficiently large N, this can only hold if $|S| \leq c_2$. □

3.2. **Genericity.** Our strategy to prove Proposition 2.2 will be to partition S into many lower dimensional subsets and apply induction. However, the main obstacle we encounter in doing so (and which is not merely a technical issue, as can be seen from the examples in §5) is the possibility that the resulting subsets are rather independent from each other, in the sense that they do not share many residue classes. If this happens, then the fact that a small polynomial vanishes at one of this subsets will not give us much information about the behavior of this polynomial in the other subsets. However, in order for this to happen it would be necessary for these subsets to occupy very few residue classes and this would imply the existence of too many elements in each subset occupying the same residue class. While with our hypothesis one cannot guarantee that this never happens, the goal of this section is to show that this indeed does not happen on average, which will be sufficient for our arguments.

We begin with the following definition.

Definition 3.3 (Genericity)**.** Given a real number $B > 0$ and some integer $l > 0$ we say that a set $S \subseteq [N]^d$ is (B, l)-generic mod p if

$$\frac{|S(\mathbf{a}; p)|}{|S|} < \frac{B}{p^l},$$

for every residue class $\mathbf{a} (\mathrm{mod}\ p)$.

Given a set of primes $P \subseteq \mathcal{P}(Q)$ we shall write $P' \hookrightarrow P$ to mean a subset $P' \subseteq P$ with $w(P') \gg w(P)$. Recall that by our conventions in §2.1 the implied constants depend on the parameters $d, h, \varepsilon, \kappa$ of Proposition 2.2. The rest of this section is devoted to the proof of the following lemma.

Lemma 3.4. *Let $d, h \geq 1$ be arbitrary integers and let $\varepsilon > 0$ be some positive real number. Set $Q = N^{\frac{\varepsilon}{2d}}$ and let $P \subseteq \mathcal{P}(Q)$ satisfy $w(P) \geq \kappa \log Q$ for some $\kappa > 0$. Suppose $S \subseteq [N]^d$ is a set of size $|S| \gg N^{d-h-1+\varepsilon}$ occupying at most αp^{d-h} residue classes mod p for every prime $p \in P$ and some $\alpha > 0$. Then there exists $B = O(1)$ and a set of primes $P' \hookrightarrow P$ such that for each $p \in P'$ there is some subset $\mathcal{G}_p(S) \subseteq S$, $|\mathcal{G}_p(S)| \gg |S|$, which is $(B, d-h)$-generic mod p.*

Remarks. Here again the exact value of Q is not important as long as it is a small power of N. Also, as in the previous statements, all the hypothesis are necessary because of the examples in §5.

Proof. From now on fix an integer $h \geq 1$. If $d \leq h$ the result is trivial, so we may proceed by induction on d. Thus, let $d \geq h+1$ be some integer and assume the result holds for every smaller dimension.

Take S and P as in the statement and recall that $\pi_i(S)$ is the projection of S to the ith coordinate. We claim that for some $1 \leq i \leq d$ there exists a set $S' \subseteq S$ with $|S'| \geq |S|/2^d$ such that every $A \subseteq S'$ with $|A| \geq |S'|/2$ satisfies $|\pi_i(A)| \geq Q$. Indeed, if the claim fails with $S' = S$ and $i = 1$ we may find some subset $S_1 \subseteq S$ with $|S_1| \geq |S|/2$ and $|\pi_1(S_1)| < Q$. Then, if the claim fails again with $S' = S_1$ and $i = 2$, we get some $S_2 \subseteq S_1$ with $|S_2| \geq |S_1|/2 \geq |S|/4$ and $|\pi_1(S_2)|, |\pi_2(S_2)| < Q$. Iterating this d times either we get the claim or end up with a set $S_d \subseteq S$ satisfying

$$|S| \leq 2^d |S_d| \leq 2^d |\pi_1(S_d)| \dots |\pi_d(S_d)| < 2^d Q^d.$$

By our choice of Q this is clearly absurd for sufficiently large N and therefore the claim follows.

Since it sufficies to prove the lemma for such a subset S' we may assume without lost of generality that $S' = S$ and permuting the coordinates if necessary we may also assume $i = 1$. Hence, we have that

$$|\pi_1(A)| \geq Q \text{ for every } A \subseteq S \text{ with } |A| \geq |S|/2. \tag{3.4}$$

We wish to construct a dense subset of S which is in an adequate position to apply the induction hypothesis. Since we will be working with the first coordinate, given some $a \in \mathbb{Z}/p\mathbb{Z}$, we will write $S(a;p)$ to refer to those elements of S having their first coordinate congruent to $a (\mathrm{mod}\ p)$. Let B_1 be some large constant to be specified later. Since $|[S]_p| \leq \alpha p^{d-h}$, it is clear that there can be at most $\alpha p/B_1$ residue classes $a \in [\pi_1(S)]_p \subseteq \mathbb{Z}/p\mathbb{Z}$ for which $|[S(a;p)]_p| \geq B_1 p^{d-h-1}$. We denote by $\mathcal{E}_1(p)$ this exceptional set. Also, we write

$$\mathcal{E}_2(p) := \left\{ a \in [\pi_1(S)]_p : |S(a;p)| \geq \frac{B_1}{\alpha p}|S| \right\}.$$

From the obvious fact that $\sum_{a \in \mathbb{Z}/p\mathbb{Z}} |S(a;p)| = |S|$ it follows that $|\mathcal{E}_2(p)| \leq \alpha p/B_1$ and therefore $|\mathcal{E}(p)| \leq 2\alpha p/B_1$, where $\mathcal{E}(p) := \mathcal{E}_1(p) \cup \mathcal{E}_2(p)$. By means of the larger sieve we may now deduce that not too many integers in $[N]$ can lie in $\mathcal{E}(p)$ for many $p \in P$. Indeed, consider the set X which consists of all elements $x \in [N]$ for which

$$\sum_{p \in P} \mathbf{1}_{x (\mathrm{mod}\ p) \in \mathcal{E}(p)} \frac{\log p}{p} \geq \frac{1}{2} w(P).$$

By the pigeonhole principle, one may then find a set of primes $P_1 \subseteq P$ with $w(P_1) \geq \frac{1}{4} w(P)$ and such that $|\bigcup_{a \in \mathcal{E}(p)} X(a;p)| \geq \frac{1}{4}|X|$ for every $p \in P_1$. It then follows from Lemma 3.1 that upon choosing B_1 sufficiently large, we can ensure that $|X| < Q$.

By (3.4), we deduce that $\left|S \setminus \pi_1^{-1}(X)\right| \geq \frac{1}{2}|S|$. We may therefore find a subset $S' \subseteq S$ with $|S'| \geq \frac{1}{4}|S|$ which does not intersect $\pi_1^{-1}(X)$ and such that $S'_x := \pi_1^{-1}(x) \cap S'$ satisfies $|S'_x| \gg N^{d-h-2+\varepsilon}$ for every $x \in \pi_1(S')$. Every such x lies outside of X and therefore has associated a set of primes $P_x \hookrightarrow P$ for which $x (\mathrm{mod}\ p) \notin \mathcal{E}(p)$. Since $\mathcal{E}_1(p) \subseteq \mathcal{E}(p)$, we may apply the induction hypothesis to S'_x for every x to see that there exists sets of primes $P'_x \hookrightarrow P_x$ and constants $c, B_2 > 0$ independent of x, such that for each $p \in P'_x$ there is a $(B_2, d-h-1)$-generic mod p set $\mathcal{G}_p(S'_x) \subseteq S'_x$ containing at least $c\,|S'_x|$ elements.

Since the sets P'_x constructed above satisfy $P'_x \hookrightarrow P$, with the implied constant independent of x, we may apply again the pigeonhole principle to locate some set of primes $P' \hookrightarrow P$ and some constant $c > 0$, such that for each $p \in P'$ there are at

least $c|S'|$ elements $s \in S'$ for which $p \in P'_{\pi_1(s)}$. It thus follows that if for a prime $p \in P'$ we consider the set

$$\mathcal{G}_p(S) := \bigcup_{x : p \in P'_x} \mathcal{G}_p(S'_x),$$

then $|\mathcal{G}_p(S)| \gg |S'| \gg |S|$ and $\mathcal{G}_p(S) \cap \pi_1^{-1}(x) = \mathcal{G}_p(S'_x)$ is a $(B_2, d-h-1)$-generic set for every $x \in \pi_1(\mathcal{G}_p(S))$. Also, we see that there are at most $\frac{B_1}{\alpha p}|S| \ll \frac{B_1}{\alpha p}|\mathcal{G}_p(S)|$ elements of $\mathcal{G}_p(S)$ having the same first coordinate mod p since by construction it does not lie in $\mathcal{E}_2(p)$. It thus follows that $\mathcal{G}_p(S)$ is a B-generic set for some large B depending on B_1 and B_2 but independent of p and this concludes the proof of Lemma 3.4. □

4. The proof of Proposition 2.2

In this section we give a proof of Proposition 2.2. As we did in the proof of Lemma 3.4 we will fix an integer h and induct on d. Since for $d \leq h$ the result is either trivial or follows from Lemma 3.2 we may assume $d \geq h+1$ and that the result holds for all smaller dimensions.

Before we proceed, we give a brief discussion of the strategy of the proof. By our size hypothesis on S, we know that there must exist some coordinate i such that both the projection $\pi_i(S)$ and the corresponding sections $\pi_i^{-1}(x) \cap S$ are big (at least on average), and therefore generically distributed in the residue classes they occupy (by Lemma 3.4). Combining these two facts, one can deduce the existence of $m \gg r$ sections $\pi_i^{-1}(x) \cap S$ of S such that the probability of some $s \in S$ of being congruent mod p to some element of these sections is roughly m/p for many p (see Lemma 4.1). This in turn implies that if an r-polynomial f vanishes at these sections, $f(s)$ is expected to be divisible by many primes, which by the boundedness of f would imply that $f(s) = 0$. Thus, it only remains to find a set that is characteristic for these m sections, but by the induction hypothesis each section admits a characteristic subset and the result then follows by taking the union of these.

We now turn to the details. To help the reader, we begin with a summary of previously introduced notation that will be needed during the proof.

- $|A|$ - the cardinality of a set A,
- $w(P) := \sum_{p \in P} \frac{\log p}{p}$.
- $P' \hookrightarrow P$ - a subset of primes $P' \subseteq P$ with $w(P') \gg w(P)$,
- $\pi_i(A)$ - the projection of A to the ith coordinate,
- $A_x := \pi_1^{-1}(x) \cap A$ for a set $A \subseteq [N]^d$ (but P_x will have a different meaning for P a set of primes),
- $[S]_p$ - the set of residue classes occupied by S mod p,
- $S(\mathbf{a}; p)$ - those elements of S congruent to $\mathbf{a}$ mod p,
- $S(a; p)$ - those elements of S with first coordinate congruent to a mod p,

We are thus given a set S and some positive integer r. Our first step will be to find generic sets inside the sections of S for many primes p. Proceeding as in the beginning of the proof of Lemma 3.4 we may assume that

$$|\pi_1(A)| \geq Q \text{ for every } A \subseteq S \text{ with } |A| \geq |S|/2. \tag{4.1}$$

This allows us, at the cost of passing to a subset of half density if necessary, to get the bound

$$|S_x| \leq 2|S|/Q \text{ for every } x \in [N], \tag{4.2}$$

where $S_x := \pi_1^{-1}(x) \cap S$. Finally we may also assume, again by passing to a subset of half density if necessary, that $|S_x| \gg N^{d-h-2+\varepsilon}$ for every $x \in \pi_1(S)$.

Let B be some large constant. For every prime p we denote by $\mathcal{E}(p)$ the set of residue classes $a \in \mathbb{Z}/p\mathbb{Z}$ for which $|[S(a;p)]_p| \geq Bp^{d-h-1}$ (recall that $S(a;p)$ stands for those elements of S having their first coordinate congruent to $a (\mathrm{mod}\ p)$ and thus $[S(a;p)]_p$ consists of those residue classes in $[S]_p \subseteq (\mathbb{Z}/p\mathbb{Z})^d$ having a as a first coordinate). Since $|\mathcal{E}(p)| \leq \alpha p/B$, applying Lemma 3.1 as in the proof Lemma 3.4, we conclude by (4.1) that if B is chosen sufficiently large, we may find some $S' \subseteq S$, $|S'| \gg |S|$, such that for each $x \in \pi_1(S')$ we have $P_x \hookrightarrow P$, with the implied constant independent of x, and where

$$P_x := \{p \in P : x (\mathrm{mod}\ p) \notin \mathcal{E}(p)\}.$$

This places us in a position in which we can apply the induction hypothesis to each section S'_x of S' to find some $\delta_0 \gg 1$ independent of x such that each S'_x admits a (r, δ_0)-characteristic subset of size $O(r^{d-h-1})$. In particular, we see that at the cost of passing to a subset of S' of density δ_0 if necessary, we may assume that inside each S'_x we can find a set of size $O(r^{d-h-1})$ which is $(r,1)$-characteristic for the whole section. Notice that since we are refining the sections, we still get a bound of the form $|S'_x| \gg N^{d-h-2+\varepsilon}$ for every $x \in \pi_1(S')$. Thus, we may also apply Lemma 3.4 to every such S'_x obtaining sets of primes $P'_x \hookrightarrow P_x$ such that for every $p \in P'_x$ we can find a $(C, d-h-1)$-generic subset $\mathcal{G}_p(S_x) \subseteq S'_x$, $|\mathcal{G}_p(S_x)| \gg |S'_x|$, where C and the implied constants are independent of p and x. In particular, we may find some set of primes $P' \hookrightarrow P$ such that for each $p \in P'$ the set

$$\mathcal{G}_p(S) := \bigcup_{x : p \in P'_x} \mathcal{G}_p(S_x),$$

satisfies $|\mathcal{G}_p(S)| \gg |S'| \gg |S|$ and each nonempty section $(\mathcal{G}_p(S))_x$ of $\mathcal{G}_p(S)$ is a $(C, d-h-1)$-generic set.

From now on we write $\mathcal{G}_p := \mathcal{G}_p(S)$. The next lemma is crucial as it allows us to find sections of S containing the residue class of many elements of S for many primes p.

Lemma 4.1. *There exists a set $\mathcal{B} \subseteq S'$, $|\mathcal{B}| \gg |S|$, such that for every non empty section $\mathcal{B}_x$ of $\mathcal{B}$ there is a set of primes $P_x \hookrightarrow P' \hookrightarrow P$ with*

$$\left|\left\{s \in S' : [s]_p \in [\mathcal{B}_x]_p\right\}\right| \geq \frac{c|S|}{p},$$

for every $p \in P_x$, where $c > 0$ does not depend on x or p.

Proof. We begin by fixing a prime $p \in P'$ and considering some residue class $a \in [\pi_1(\mathcal{G}_p)]_p$. Since p is fixed we will simply write $\mathcal{G}_p(a)$ to denote those elements of $\mathcal{G}_p$ with first coordinate congruent to $a (\mathrm{mod}\ p)$. Also, given a class $\mathbf{b} \in (\mathbb{Z}/p\mathbb{Z})^d$ we write $\mathcal{G}_p(\mathbf{b})$ for those elements of $\mathcal{G}_p$ congruent to $\mathbf{b} (\mathrm{mod}\ p)$. By the pigeonhole principle and the fact that by construction of P' it is $|\,[\mathcal{G}_p(a)]_p\,| \leq Bp^{d-h-1}$ it follows that we may find some $\mathbf{b}_1 \in [\mathcal{G}_p(a)]_p \subseteq (\mathbb{Z}/p\mathbb{Z})^d$ with

$$|\mathcal{G}_p(\mathbf{b}_1)| \geq |\mathcal{G}_p(a)|/(Bp^{d-h-1}).$$

Consider now the set $\mathcal{B}_1 \subseteq \mathcal{G}_p(a)$ defined by

$$\mathcal{B}_1 := \bigcup_{s : [s]_p = \mathbf{b}_1} (\mathcal{G}_p)_{\pi_1(s)}, \tag{4.3}$$

that is, $\mathcal{B}_1$ is the union of those sections $(\mathcal{G}_p)_x$ in $\mathcal{G}_p$ containing a representative of $\mathbf{b}_1$.

Since each $(\mathcal{G}_p)_x$ is a $(C, d-h-1)$-generic set, we have that

$$|(\mathcal{G}_p)_x| \geq \frac{p^{d-h-1}}{C} |(\mathcal{G}_p)_x(\mathbf{b}_1)|$$

and therefore

$$|\mathcal{B}_1| \geq \frac{p^{d-h-1}}{C}\,|\mathcal{G}_p(\mathbf{b}_1)| \geq \frac{1}{BC}|\mathcal{G}_p(a)|. \tag{4.4}$$

Notice now that since $|\mathcal{G}_p(a)| \geq |\mathcal{B}_1|$ and $|[\mathcal{G}_p(a)]_p| \leq Bp^{d-h-1}$, by the first inequality of (4.4) and the pigeonhole principle we may find another residue class $\mathbf{b}_2 \in [\mathcal{G}_p(a)]_p$ with

$$\begin{aligned}|\mathcal{G}_p(\mathbf{b}_2)| &\geq \frac{1}{Bp^{d-h-1}}|\mathcal{G}_p(a) \setminus \mathcal{G}_p(\mathbf{b}_1)| \\ &\geq \frac{1}{Bp^{d-h-1}}\left(1 - \frac{C}{p^{d-h-1}}\right)|\mathcal{G}_p(a)|,\end{aligned}$$

which is at least $|\mathcal{G}_p(a)|/(2Bp^{d-h-1})$ if $p^{d-h-1} > 2C$. In such a case, if we now define $\mathcal{B}_2$ as in (4.3), but this time with respect to $\mathbf{b}_2$, the same reasoning that gives (4.4) implies $|\mathcal{B}_2| \geq \frac{1}{2BC}|\mathcal{G}_p(a)|$. Iterating this process we end up with a sequence $\mathbf{b} = \{\mathbf{b}_1, \dots, \mathbf{b}_q\}$ of residue classes, $q = \lceil \frac{p^{d-h-1}}{2C} \rceil$, satisfying

$$\begin{aligned}|\mathcal{G}_p(\mathbf{b}_j)| &\geq \frac{1}{Bp^{d-h-1}}\left|\mathcal{G}_p(a) \setminus \bigcup_{i=1}^{j-1}\mathcal{G}_p(\mathbf{b}_i)\right| \\ &\geq \frac{1}{Bp^{d-h-1}}\left(1 - \frac{(q-1)C}{p^{d-h-1}}\right)|\mathcal{G}_p(a)| \\ &\geq \frac{|\mathcal{G}_p(a)|}{2Bp^{d-h-1}},\end{aligned}$$

and $|\mathcal{B}_j| \geq \frac{1}{2BC}|\mathcal{G}_p(a)|$. In particular, we have that

$$\sum_{j=1}^{q}|\mathcal{B}_j| \geq \frac{q}{2BC}|\mathcal{G}_p(a)|. \tag{4.5}$$

Now, we consider the set

$$\mathcal{B}[a] := \left\{ s \in \mathcal{G}_p(a) : \sum_{j=1}^{q}\mathbf{1}_{s\in\mathcal{B}_j} \geq \frac{q}{4BC} \right\}.$$

Notice that $\mathcal{B}[a]_x := \mathcal{B}[a] \cap \pi_1^{-1}(x)$ equals $(\mathcal{G}_p)_x$ whenever this intersection is not empty. Also, (4.5) implies

$$|\mathcal{B}[a]| \geq \frac{1}{4BC}|\mathcal{G}_p(a)|. \tag{4.6}$$

We see that $\mathcal{B}[a]$ is very close to what we want, since if we take any nonempty section $\mathcal{B}[a]_x$ of this set, then there are at least $|\mathcal{G}_p(a)|/(4BC)^2$ elements $s \in \mathcal{G}(a)$ such that $s \equiv y (\mathrm{mod}\ p)$ for some $y \in \mathcal{B}[a]_x$.

We now let $\mathcal{R} \subseteq [\pi_1(S)]_p$ consist of those residue classes $a \in \mathbb{Z}/p\mathbb{Z}$ with $|\mathcal{G}_p(a)| \geq \frac{1}{2p}|\mathcal{G}_p|$ and write

$$\mathcal{B}[p] := \left\{ s \in S' : S'_{\pi_1(s)} \cap \mathcal{B}[a] \neq \emptyset \text{ for some } a \in \mathcal{R} \right\}.$$

In other words, $\mathcal{B}[p]$ consists of those sections of S' intersecting $\bigcup_{a\in\mathcal{R}}\mathcal{B}[a]$. In particular, since $\mathcal{B}[p]$ contains the disjoint union $\bigcup_{a\in\mathcal{R}}\mathcal{B}[a]$, we see from (4.6) and the definition of $\mathcal{R}$ that

$$|\mathcal{B}[p]| \geq \frac{1}{8BC}|\mathcal{G}_p| \geq c|S|,$$

for some constant c independent of p.

Recall now that $w(P') \geq c \log Q$. For an element $s \in S'$ write P'_s for the set of primes $p \in P'$ for which $s \in \mathcal{B}[p]$. It follows from the above paragraph that for an appropriate choice of c the set

$$\mathcal{B} := \{s \in S' : w(P'_s) \geq c \log Q\}, \tag{4.7}$$

satisfies $|\mathcal{B}| \geq c|S|$. It is easy to check that $\mathcal{B}$ is of the desired form. □

To conclude the proof of Proposition 2.2 we will show that if an r-polynomial vanishes at the sections $\mathcal{B}_x$ for $\gg_r 1$ distinct values of x, then it must also vanish at a positive proportion of S. To this end, we choose m distinct sections of S' having nontrivial intersection with $\mathcal{B}$, where $m = O_r(1)$ is to be specified later. Notice that by (4.2) and Lemma 4.1 this is always possible provided N is sufficiently large. Call $\mathcal{L} := S'_{x_1} \cup \ldots \cup S'_{x_m}$ the union of these sections. Let $P_{\mathcal{L}}$ consist of those primes p for which there exists a pair of sections $S'_{x_i} \neq S'_{x_j}$ in $\mathcal{L}$ with $[S'_{x_i}]_p \cap [S'_{x_j}]_p \neq \emptyset$. Given such a pair of sections the fact that $[S'_{x_i}]_p \cap [S'_{x_j}]_p \neq \emptyset$ implies in particular that $x_i \equiv x_j (\mathrm{mod}\ p)$. Since $x_i \neq x_j$ this implies that the sum of $\log p$ over such primes is bounded by $\log N$. Thus, we see that

$$\sum_{p \in P_{\mathcal{L}}} \log p \leq \binom{m}{2} \log N, \tag{4.8}$$

and this implies that $w(P_{\mathcal{L}}) \ll_r \log \log N$.

We now consider on S' the function

$$\psi_{\mathcal{L}}(s) := \sum_{p \leq Q} \mathbf{1}_{\exists x \in \mathcal{L} : s \equiv x (\mathrm{mod}\ p)} \log p.$$

Thus, $\psi_{\mathcal{L}}(s)$ measures the extent to which the residue classes occupied by s have a representative in $\mathcal{L}$. If we write P_i to denote the set of primes in Lemma 4.1 corresponding to the section $S'_{x_i} \cap \mathcal{B}$ of $\mathcal{B}$, it follows from this lemma and (4.8) that

$$\begin{aligned}
\sum_{s \in S'} \psi_{\mathcal{L}}(s) &\geq \sum_{i=1}^{m} \sum_{p \in P_i \setminus P_{\mathcal{L}}} \sum_{s \in S'} \mathbf{1}_{\exists x \in S'_{x_i} : s \equiv x (\mathrm{mod}\ p)} \log p \\
&\geq \sum_{i=1}^{m} \sum_{p \in P_i \setminus P_{\mathcal{L}}} \frac{c|S|}{p} \log p \\
&\geq m|S| \left(c \log Q + O_r(\log \log N)\right) \\
&\geq c_0 m |S| \log Q,
\end{aligned}$$

for some $c_0 > 0$ and sufficiently large N.

Set $\delta = \frac{\varepsilon c_0}{4d}$ and suppose there are at most $\delta|S|$ elements $s \in S'$ with $\psi_{\mathcal{L}}(s) \geq 3r \log N$. Since $\psi_{\mathcal{L}}(s) \leq m \log N$ for every $s \notin \mathcal{L}$ we conclude that

$$c_0 m |S| \log Q \leq |\mathcal{L}| 2Q + |S| 3r \log N + \delta |S| m \log N,$$

where we used that $\sum_{p \leq Q} \log p \leq 2Q$ for large Q. Hence, by (4.2) we derive that

$$m \left(\frac{\varepsilon c_0}{2d} - \delta - \frac{4}{\log N} \right) \leq 3r.$$

Taking $m = 7r/\delta$ we get a contradiction for sufficiently large N. We may therefore assume that the set

$$A := \{s \in S' : \psi_{\mathcal{L}}(s) \geq 3r \log N\},$$

has size $|A| \geq \delta |S|$ for the above choices of m and δ.

We will now show that if an r-polynomial vanishes at $\mathcal{L}$, then it also vanishes at A. Indeed, let f be such a polynomial and let $x \in A$ be arbitrary. By definition, we

have $|f(x)| < N^{3r}$. On the other hand, if p is a prime for which there exists some $y \in \mathcal{L}$ with $x \equiv y (\mathrm{mod}\ p)$, then the fact that $f(y) = 0$ implies that $p|f(x)$. But by definition of A the product of all such p is at least N^{3r} so we see that the only way for this to hold is to have $f(x) = 0$, which proves our claim.

By the induction hypothesis and our construction of S' we know that for each $S'_{x_i} \in \mathcal{L}$ we may find a $(r, 1)$-characteristic set of size $O(r^{d-h-1})$. Taking the union of these m sets we have thus found a set of size $O(r^{d-h})$ which is (r, δ)-characteristic for S, with δ as above. This concludes the proof of Proposition 2.2.

5. Ill-distributed sets with no algebraic structure

In this section we provide some examples of high dimensional ill-distributed sets possessing no algebraic structure. In particular, we show that the assertion of Theorem 1.1 fails when $\varepsilon = 0$. To begin with, we use a slight modification of the construction given in [11, §4.3] to see that, given any $0 < \eta < 1$, one may construct a subset of $[N]$ of size $\gg (\log N)^\eta$ which occupies at most p^η residue classes for every prime p and which possesses no algebraic structure. Indeed, if N is sufficiently large, we may find some integer Q with $Q < \log N < 2Q$ such that the product of all primes $p \leq Q$, say R, satisfies $N^{1/4} < R < N$ (this, of course, is very crude). For each prime $p \leq Q$ choose $\lfloor p^\eta \rfloor$ residue classes. Then, by the Chinese remainder theorem, there are $\sim R^\eta$ elements below R belonging to a selected class for every $p \leq Q$. Choose $\lfloor (\log N)^\eta / 2 \rfloor$ of these elements and call this set X. Notice that for all primes $p > Q$ we have $p^\eta > |X|$ and therefore X occupies at most p^η residue classes for these primes p. Since by construction it also occupies that many classes for all primes $p \leq Q$, we get the claim.

We now proceed to give some examples of ill-distributed sets with no algebraic structure. The first one already shows that Theorem 1.1 is best possible.

Example 5.1. This follows readily from the above construction. Fix some pair of positive integers d, h with $d \geq h + 1$ and consider $h + 1$ different sets $X_1, \ldots, X_{h+1}$ constructed as in the previous paragraph with $\eta = 1/(h + 1)$. If we define the set

$$S := \left\{ (x_1, \ldots, x_d) \in [N]^d : x_i \in X_i \ \forall 1 \leq i \leq h + 1 \right\},$$

then we have that $|S| \gg N^{d-h-1} \log N$ while $|[S]_p| \leq p^{d-h}$ for every prime p, from where it follows that we cannot take $\varepsilon = 0$ in Theorem 1.1.

Example 5.2. One can generalize the above example by "perturbing" arbitrary algebraic sets. We show a simple instance of this. Let $d = 3$ and consider two polynomials $f, g \in \mathbb{Z}[x]$. Let X and Y be sets of size $\gg (\log N)^{1/2}$ occupying at most $p^{1/2}$ residue classes for every prime p. Then, we see that

$$\{(x, f(x) \cdot X, g(x) \cdot Y) : x \in [N]\}$$

is a big set of integer points occupying at most p^2 residue classes.

Finally, we show that not all counterexamples are perturbations of strongly algebraic sets.

Example 5.3. Fix some small $\varepsilon > 0$. By the Chinese remainder theorem one can construct a set $X \subseteq [N]$ of size $|X| \sim N^{1-\varepsilon}$ occupying only one residue class for every prime $p \leq \varepsilon \log N$. Take $K = \lfloor (\varepsilon \log N)^{1/3} \rfloor$ and let $f_1, \ldots, f_K, g_1, \ldots, g_K$ be a family of polynomials. Also, let $X_1, \ldots, X_K, Y_1, \ldots, Y_K$ be arbitrary sets of size at most $(\varepsilon \log N)^{1/3}$. Then

$$\bigcup_{i=1}^{K} \{(x, f_i(x) \cdot X_i, g_i(x) \cdot Y_i) : x \in X\}$$

is a big set of integer points occupying at most p^2 residue classes for every prime p. Notice that this construction is of a different nature than the one given in Example 5.2, since the union of that many algebraic sets may not retain any algebraic structure itself.

It follows from the above examples that strange things can happen if one allows the set to possess too many very small sections. However, we shall show in Theorem 6.1 below that the methods of this paper do indeed work as long as one avoids this type of situations.

6. Further results and conjectures

6.1. **A generalization of Theorem 1.1.** We now state the most general result which follows at once from the methods of this paper. Let $0 \leq k < d$ be integers and let $\varepsilon > 0$ be some positive real number. We say a set $S \subseteq [N]^d$ is $(1, \varepsilon)$-regular if $|S| \geq N^{\varepsilon}$. Recursively, we say $S \subseteq [N]^d$ is (k, ε)-regular if there exists some $1 \leq i \leq d$ such that for every $x \in [N]$

(1) $|\pi_i^{-1}(x) \cap S| \leq |S|/N^{\varepsilon}$,
(2) $\pi_i^{-1}(x) \cap S$ is either empty or $(k-1, \varepsilon)$-regular.

The first condition allows us to recover (4.2), while the second one enables us to use Lemma 3.4 and the induction hypothesis as it was done in the main argument. As a consequence, one recovers the conclusions preceding Lemma 4.1 and from here the proof of (the analogous of) Proposition 2.2 proceeds without further modifications. One can thus deduce that any (k, ε)-regular S set occupying $\ll p^k$ residue classes admits a bounded polynomial vanishing at a positive proportion of S. Furthermore, since it is easy to see that any subset $S' \subseteq S$ of a (k, ε)-regular set with $|S'| \geq \eta |S|$ admits a $(k, \varepsilon/2)$-regular subset of half density (provided N is sufficiently large in terms of η) we can in fact deduce the following stronger result.

Theorem 6.1. *Let $0 \leq k < d$ be integers and let $\varepsilon, \eta, \alpha$ be positive real numbers. Then there exists $C = O_{\varepsilon,\eta,\alpha,k,d}(1)$ such that for every (k, ε)-regular set $S \subseteq [N]^d$ occupying less than αp^k residue classes for every prime p, there exists a polynomial $f \in \mathbb{Z}[x_1, \ldots, x_d]$ of degree at most C and coefficients bounded by N^C vanishing at more than $(1-\eta)|S|$ points of S.*

One would expect a reasonable set to be well approximated by a bounded union of (k, ε)-regular subsets, in which case it is clear that the same conclusion holds. For example, it was implicitly shown in the proof of Theorem 1.1 that given any $\eta > 0$ and any $S \subseteq [N]^d$ with $|S| \gg N^{k-1+\varepsilon}$ there exists some $S' \subseteq S$ with $|S'| \geq (1-\eta)|S|$ which is the union of a bounded number (in terms of η and d) of $(k, \varepsilon/4d)$-regular subsets.

Finally, it is important to note that one cannot hope to do much better than Theorem 6.1 in this generality, since the regularity conditions are necessary in order to avoid those constructions emerging from the Chinese Remainder Theorem as in §5.

6.2. **Approximate reduction.** We shall give a quick application of Theorem 6.1 to the study of functions preserving some structure when reduced modulo a prime, that is, functions f for which knowing the class of $x (\mathrm{mod}\ p)$ gives us information about the class of $f(x) (\mathrm{mod}\ p)$. Thus, given a positive integer K, we say a function $f : [N]^k \to [N^r]^t$ has K-approximate reduction if $\left| \left[f\left([N]^k(\mathbf{a}) \right) \right]_p \right| \leq K$ for every $\mathbf{a} \in (\mathbb{Z}/p\mathbb{Z})^k$ and every prime p. When $K = 1$ this implies the very strong property of *recurrence* mod p and using this, it was shown by Hall [9] and Ruzsa [15] (see also [16, §XV.41]) that for large N the only functions having 1-approximate reduction

are polynomials (notice that we are assuming our functions to have polynomial growth, which is in fact a necessary condition [9]). Since the graph of a function $f : [N]^k \to [N^r]^t$ is always a $(k, 1/2r)$-regular set, it follows from Theorem 6.1 that this is indeed a very robust phenomenon:

Corollary 6.2. *Suppose $f : [N]^k \to [N^r]^t$ has K-approximate reduction and let $\Gamma(f)$ be the graph of f. Then there exists $C = O_{k,r,t,K}(1)$ and a polynomial $P \in \mathbb{Z}[x_1, \dots, x_d]$ of degree at most C and coefficients bounded by N^C, such that P vanishes at more than $(1-\eta)|\Gamma(f)|$ points of $\Gamma(f)$.*

6.3. **The Inverse Sieve Problem for $d = 1$.** We conclude by mentioning a very strong version of the inverse sieve problem which is conjectured to hold for sets $S \subseteq [N]$ (see [3, Problem 7.2] and [11]).

Conjecture 6.3. *Suppose $S \subseteq [N]$ is some set of integers of size $|S| \geq N^\varepsilon$ occupying less than αp residue classes for some $0 < \alpha < 1$ and every prime p. Then most of S is contained in the image of an integer polynomial of degree bounded in terms of α and ε.*

As a more precise instance of this, they conjecture for example that if a set S has size $|S| \geq N^{0.49}$ say, and occupies less than $2p/3$ residue classes mod p for every prime p, then most of S must be contained in a set of the form $\{an^2 + bn + c : n \in \mathbb{Z}\}$. This can be seen as an inverse conjecture for the large sieve [5, 14].

Conjecture 6.3 seems to be hard. For example, it was shown by Green that if the residue classes occupied by S lie outside some interval of length $(p-1)/2$ then $|S| \ll_\varepsilon N^{1/3+\varepsilon}$ for any $\varepsilon > 0$. However, even in this particular case, to get a bound of the form $|S| \ll_\varepsilon N^\varepsilon$, it seems necessary to appeal to very deep conjectures of analytic number theory like the exponent pair conjecture (see [5]). Furthermore, as noted by Helfgott and Venkatesh [11, §4.2], Conjecture 6.3 implies that there are $\ll_\varepsilon N^\varepsilon$ points on an irrational curve within a square of side N, which is itself a well known open problem.

References

[1] E. Bombieri and J. Pila, *The number of integral points on arcs and ovals*, Duke Math. J. **59** (1989), 337-357.

[2] A. C. Cojocaru and M. R. Murty, *Introduction to Sieve Methods and Their Applications*. London Mathematical Society Student Texts, Cambridge University Press, 2005.

[3] E. Croot and V. F. Lev, *Open problems in additive combinatorics*, in Additive Combinatorics, CRM Proc. Lecture Notes **43**, 207-233, Amer. Math. Soc., Providence, RI, 2007.

[4] B. J. Green, *Approximate groups and their applications: work of Bourgain, Gamburd, Helfgott and Sarnak*, Current Events Bulletin of the AMS, 2010.

[5] B. J. Green, *On a variant of the large sieve*, preprint, `arXiv:0807.5037`.

[6] B. J. Green and I. Z. Ruzsa, *Freiman's theorem in an arbitrary abelian group*, J. Lond. Math. Soc. (2) **75** (2007), no. 1, 163-175.

[7] B. J. Green, T. Tao and T. Ziegler, *An inverse theorem for the Gowers $U^{s+1}[N]$-norm*, preprint, `arXiv:1009.3998`

[8] P. X. Gallagher, *A larger sieve*, Acta Arith. **18** (1971), 77-81.

[9] R. R. Hall, *On pseudopolynomials*, Mathematika **18** (1971), 71-77.

[10] H. A. Helfgott, *Growth and generation in $SL_2(\mathbb{Z}/p\mathbb{Z})$*, Ann. of Math. (2) **167** (2008), no. 2, 601-623.

[11] H. A. Helfgott and A. Venkatesh, *How small must ill-distributed sets be?*, Analytic number theory. Essays in honour of Klaus Roth. Cambridge University Press, 2009, 224-234.

[12] E. Kowalski, *The ubiquity of surjective reduction in random groups*, notes available at `http://www.math.ethz.ch/~kowalski/notes-unpublished.html`.

[13] E. Kowalski, *The large sieve and its applications: arithmetic geometry, random walks and discrete groups*. Cambridge Tracts in Math. 175, Cambridge University Press, 2008.

[14] H. L. Montgomery, *A note on the large sieve*, J. London Math. Soc. **43** (1968), 93–98.
[15] I. Z. Ruzsa, *On congruence preserving functions*, Mat. Lapok. **22** (1971), 125-134.
[16] J. Sándor, D. S. Mitrinovic and B. Crstici, *Handbook of number theory I*, Springer, 2006.
[17] T. Tao, *Freiman's theorem for solvable groups*, Contrib. Discrete Math. **5** (2010), no. 2, 137-184.
[18] T. Tao and V. Vu, *Additive combinatorics*, Cambridge Studies in Advanced Mathematics, 105. Cambridge University Press, Cambridge, 2006.
[19] T. Tao and V. Vu, *Inverse Littlewood-Offord theorems and the condition number of random matrices*, Ann. of Math. (2) **169** (2009), 595-632.

Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina

E-mail address: mwalsh@dm.uba.ar