THE GREEN–TAO THEOREM: AN EXPOSITION

DAVID CONLON, JACOB FOX, AND YUFEI ZHAO

Abstract. The celebrated Green-Tao theorem states that the prime numbers contain arbitrarily long arithmetic progressions. We give an exposition of the proof, incorporating several simplifications that have been discovered since the original paper.

1. Introduction

In 2004, Ben Green and Terence Tao [23] proved the following celebrated theorem, resolving a folklore conjecture about prime numbers.

Theorem 1.1 (Green-Tao)**.** *The prime numbers contain arbitrarily long arithmetic progressions.*

Our intention is to give a complete proof of this theorem. Although there have been numerous other expositions [21, 22, 28, 29, 43, 45, 47], we were prompted to write this note because of our recent work [7, 51] simplifying one of the key technical ingredients in the proof. Together with work of Gowers [21], Reingold, Trevisan, Tulisiani, and Vadhan [33], and Tao [42], there have now been substantial simplifications to almost every aspect of the proof. We have chosen to collect these simplifications and present an up-to-date exposition in order to make the proof more accessible.

A key element in the proof of Theorem 1.1 is Szemerédi's theorem [41] on arithmetic progressions in dense subsets of the integers. To state this theorem, we define the *upper density* of a set $A \subseteq \mathbb{N}$ to be

$$\limsup_{N\to\infty} \frac{|A \cap [N]|}{N}, \qquad \text{where } [N] := \{1, 2, \dots, N\}.$$

Theorem 1.2 (Szemerédi)**.** *Every subset of* $\mathbb{N}$ *with positive upper density contains arbitrarily long arithmetic progressions.*

Szemerédi's theorem is a deep and important result and the original proof [41] is long and complex. It has had a huge impact on the subsequent development of combinatorics and, in particular, was responsible for the introduction of the *regularity lemma*, now a cornerstone of modern combinatorics. Numerous different proofs of Szemerédi's theorem have since been discovered and all of them have introduced important new ideas that grew into active areas of research. The three main modern approaches to Szemerédi's theorem are by ergodic theory [13, 15], higher order Fourier analysis [17, 18], and hypergraph regularity [20, 31, 34, 35, 46]. However, none of these approaches are easy. We shall therefore assume Szemerédi's theorem as a black box and explain how to derive the Green-Tao theorem using it.

As the set of primes has density zero, Szemerédi's theorem does not immediately imply the Green-Tao theorem. Nevertheless, Erdős famously conjectured that the density of the primes alone should guarantee the existence of long APs.[1] Specifically, he conjectured that any subset A of $\mathbb{N}$ with divergent harmonic sum, i.e., $\sum_{a\in A} 1/a = \infty$, must contain arbitrarily long APs. This

The first author was supported by a Royal Society University Research Fellowship.

The second author was supported by a Packard Fellowship, NSF Career Award DMS-1352121, an Alfred P. Sloan Fellowship, and an MIT NEC Corporation Award.

The third author was supported by a Microsoft Research PhD Fellowship.

[1]For brevity, we will usually write AP for arithmetic progression and k-AP for a k-term AP.

conjecture is widely believed to be true, but it has yet to be proved even in the case of 3-term APs.[2]

If not by density considerations, how do Green and Tao prove their theorem? The answer is that they treat Szemerédi's theorem as a black box and show, through a *transference principle*, that a Szemerédi-type statement holds relative to sparse pseudorandom subsets of the integers, where a set is said to be *pseudorandom* if it resembles a random set of similar density in terms of certain statistics or properties. We refer to such a statement as a *relative Szemerédi theorem*. Given two sets A and S with $A \subseteq S$, we define the relative upper density of A in S to be $\limsup_{N\to\infty} |A \cap [N]| \,/\, |S \cap [N]|$.

Relative Szemerédi theorem. (Informally) *If S is a (sparse) set of integers satisfying certain pseudorandomness conditions and A is a subset of S with positive relative density, then A contains arbitrarily long APs.*

To prove the Green-Tao theorem, it then suffices to show that there is a set of "almost primes" containing, but not much larger than, the primes which satisfies the required pseudorandomness conditions. In the work of Green and Tao, there are two such conditions, known as the linear forms condition and the correlation condition.

The proof of the Green-Tao theorem therefore falls into two parts, the first part being the proof of the relative Szemerédi theorem and the second part being the construction of an appropriately pseudorandom superset of the primes. Green and Tao credit the contemporary work of Goldston and Yıldırım [16] for the construction and estimates used in the second half of the proof. Here we will follow a simpler approach discovered by Tao [42].

The proof of the relative Szemerédi theorem also splits into two parts, the dense model theorem and the counting lemma. Roughly speaking, the dense model theorem allows us to say that if S is a sufficiently pseudorandom set then any relatively dense subset A of S may be "approximated" by a dense subset $\widetilde{A}$ of $\mathbb{N}$, while the counting lemma shows that the number of arithmetic progressions in A is close, up to a normalization factor, to the number of arithmetic progressions in $\widetilde{A}$. Since $\widetilde{A}$ is a dense subset of $\mathbb{N}$, Szemerédi's theorem implies that $\widetilde{A}$ contains arbitrarily long APs and this in turn implies that A contains arbitrarily long APs.

This is also the outline we will follow in this paper, though for each part we will follow a different approach to the original paper. For the counting lemma, we will follow the recent approach taken by the authors in [7]. This approach has significant advantages over the original method of Green and Tao, not least of which is that a weakening of the linear forms condition is sufficient for the relative Szemerédi theorem to hold. This means that the estimates involved in verifying the correlation condition may now be omitted from the proof.

In [7], the dense model theorem was replaced with a certain sparse regularity lemma. However, as subsequently observed by Zhao [51], the original dense model theorem may also be used. To prove the dense model theorem, we will follow an elegant method developed independently by Gowers [21] and by Reingold, Trevisan, Tulsiani, and Vadhan [33].

The 3-AP case of Szemerédi's theorem was first proved by Roth [36] in the 1950s. While Roth's theorem, as this case is usually known, is already a very interesting and nontrivial result, the 3-AP case is substantially easier than the general result. In contrast, when proving a relative Szemerédi theorem by *transferring* Szemerédi's theorem down to the sparse setting, the general case is not

[2]A recent result of Sanders [38] is within a hair's breadth of verifying Erdős' conjecture for 3-APs. Sanders proved that every 3-AP-free subset of $[N]$ has size at most $O(N(\log\log N)^6/\log N)$, which is just slightly shy of the logarithmic density barrier that one wishes to cross (see Bloom [4] for a recent improvement). In the other direction, Behrend [2] constructed a 3-AP-free subset of $[N]$ of size $Ne^{-O(\sqrt{\log N})}$. There is some evidence to suggest that Behrend's lower bound is closer to the truth (see [39]). For longer APs, the gap is much larger. The best upper bound, due to Gowers [18], is that every k-AP-free subset of $[N]$ has size at most $N/(\log\log N)^{c_k}$ for some $c_k > 0$ (though for $k = 4$ there have been some improvements [24]).

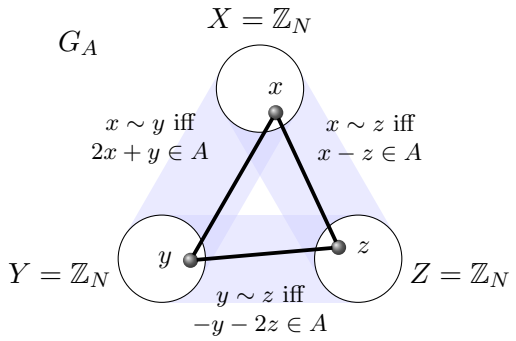


FIGURE 1. The construction in the proof of Roth's theorem.

mathematically more difficult than the 3-AP case. However, as one might expect, the notation for the general case can be rather cumbersome. For this reason, we explain various aspects of the proof first for 3-APs and only afterwards discuss how it can be adapted to the general case.

We begin, in Section 2, by presenting the Ruzsa-Szemerédi graph-theoretic approach to Roth's theorem. In particular, we present a graph-theoretic construction that will motivate the definition of the linear forms conditions, which we state in Sections 3 and 4, first for Roth's theorem, then for Szemerédi's theorem. The dense model theorem and the counting lemma are explained in Sections 5 and 6, respectively. We conclude the proof of the relative Szemerédi theorem in Section 7. In Sections 8 and 9, we will construct the relevant set of almost primes (or rather a majorizing measure for the primes) and show that it satisfies the linear forms condition. We conclude with some remarks about extensions of the Green-Tao theorem.

2. Roth's theorem via graph theory

One way to state Szemerédi's theorem is that for every fixed k every k-AP-free subset of $[N]$ has $o(N)$ elements. It is not hard to prove that this "finitary" version of Szemerédi's theorem is equivalent to the "infinitary" version stated as Theorem 1.2.

In fact, it will be more convenient to work in the setting of the abelian group $\mathbb{Z}_N := \mathbb{Z}/N\mathbb{Z}$ as opposed to $[N]$. These two settings are roughly equivalent for studying k-APs, with the only difference being that $\mathbb{Z}_N$ allows APs to wrap around 0. For example, $N-1, 0, 1$ is a 3-AP in $\mathbb{Z}_N$, but not in $[N]$. To deal with this issue, one simply embeds $[N]$ into a slightly larger cyclic group so that no k-APs wrap around zero. Working in $\mathbb{Z}_N$, we will now show how Roth's theorem follows from a result in graph theory.

Theorem 2.1 (Roth)**.** *If $A \subseteq \mathbb{Z}_N$ is 3-AP-free, then $|A| = o(N)$.*

Consider the following graph construction (see Figure 1). Given $A \subseteq \mathbb{Z}_N$, we construct a tripartite graph G_A whose vertex sets are X, Y, and Z, each with N vertices labeled by elements of $\mathbb{Z}_N$. The edges are constructed as follows (one may think of this as a variant of the Cayley graph for $\mathbb{Z}_N$ generated by A):

- $(x, y) \in X \times Y$ is an edge if and only if $2x + y \in A$;
- $(x, z) \in X \times Z$ is an edge if and only if $x - z \in A$;
- $(y, z) \in Y \times Z$ is an edge if and only if $-y - 2z \in A$.

Observe that $(x, y, z) \in X \times Y \times Z$ forms a triangle if and only if all three of

$$2x + y, \quad x - z, \quad -y - 2z$$

are in A. These numbers form a 3-AP with common difference $-x - y - z$, so we see that triangles in G_A correspond to 3-APs in A.

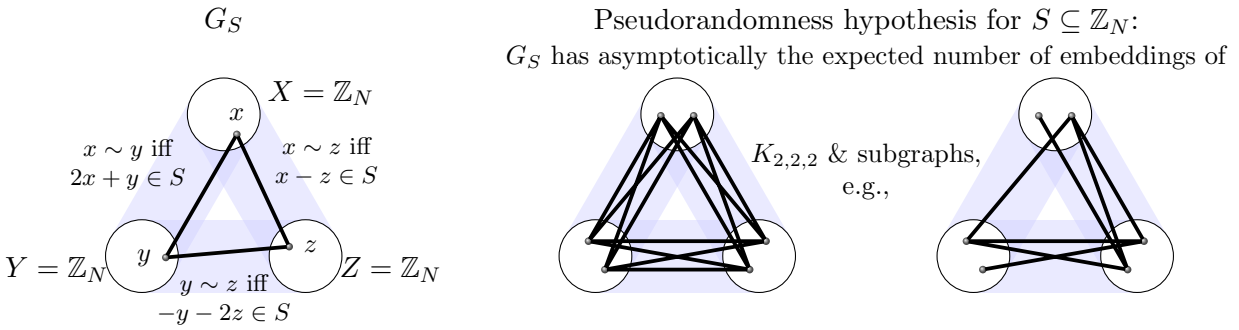


FIGURE 2. Pseudorandomness conditions for the relative Roth theorem.

However, we assumed that A is 3-AP-free. Does this mean that G_A has no triangles? Not quite. There are still some triangles in G_A, namely those that correspond to trivial 3-APs in A, i.e., 3-APs with common difference zero. So the triangles in G_A are precisely those with $x + y + z = 0$. This easily implies that every edge in G_A is contained in exactly one triangle, namely the one that completes the equation $x + y + z = 0$.

What can we say about a graph where every edge is contained in exactly one triangle? The following result of Ruzsa and Szemerédi [37] shows that it cannot have many edges.

Theorem 2.2 (Ruzsa-Szemerédi)**.** *If G is a graph on n vertices with every edge in exactly one triangle, then G has $o(n^2)$ edges.*

Our graph G_A has $3N$ vertices and $3N|A|$ edges (for every $x \in X$, there are exactly $|A|$ vertices $y \in Y$ with $2x + y \in A$ and similarly for $Y \times Z$ and $X \times Z$). So it follows by Theorem 2.2 that $3N|A| = o((3N)^2)$. Hence $|A| = o(N)$, proving Roth's theorem.

Theorem 2.2 easily follows from a result known as the *triangle removal lemma*, which says that if a graph on n vertices has $o(n^3)$ triangles, then it can be made triangle-free by removing $o(n^2)$ edges. Though both results look rather innocent, it is only recently [6, 10] that a proof was found which avoids the use of Szemerédi's regularity lemma.

We will not include a proof of Theorem 2.2 here, since this would lead us too far down the route of proving Szemerédi's theorem. However, if our purpose was not to prove Roth's theorem, then why translate it into graph-theoretic language in the first place? The reason is that the counting lemma and pseudorandomness conditions used for transferring Roth's theorem to the sparse setting are most naturally phrased in terms of graph theory.[3] We will begin to make this explicit in the next section.

3. Relative Roth theorem

In this section, we describe the relative Roth theorem. We first give an informal statement.

Relative Roth Theorem. (Informally) *If $S \subseteq \mathbb{Z}_N$ satisfies certain pseudorandomness conditions and $A \subseteq S$ is 3-AP-free, then $|A| = o(|S|)$.*

To state the pseudorandomness conditions, let $p = |S|/N$ (which may decrease as a function of N) and consider the graph G_S. This is similar to G_A, except that $(x, y) \in X \times Y$ is now made an edge if and only if $2x + y \in S$, etc. The pseudorandomness hypothesis now asks that the number of embeddings of $K_{2,2,2}$ in G_S (i.e., the number of tuples $(x_1, x_2, y_1, y_2, z_1, z_2) \in X \times X \times Y \times Y \times Z \times Z$

[3]However, it is worth stressing that the bounds in the relative Roth theorem do not reflect the poor bounds given by the graph theoretic approach to Roth's theorem. While graph theory is a convenient language for phrasing the transference principle, Roth's theorem itself only appears as a black box and any bounds we have for this theorem transfer directly to the sparse version.

FIGURE 3. The 2-blow-up of a graph is constructed by duplicating each vertex.

where x_iy_j, x_iz_j, y_iz_j are all edges for all $i, j \in \{1, 2\}$) be equal to $(1 + o(1))p^{12}N^6$, where $o(1)$ indicates a quantity that tends to zero as N tends to infinity. This is asymptotically the same as the expected number of embeddings of $K_{2,2,2}$ in a random tripartite graph of density p or in the graph G_S formed from a random set S of density p. Assuming that p does not decrease too rapidly with N, it is possible to show that with high probability the true $K_{2,2,2}$-count in these random graphs is asymptotic to this expectation. It is therefore appropriate to think of our condition as a type of pseudorandomness.

For technical reasons, it is necessary to assume that this property of having the "correct" count holds not only for $K_{2,2,2}$ but also for every subgraph H of $K_{2,2,2}$. That is, we ask that the number of embeddings of H into G_S (with vertices of H mapped into their assigned parts) be equal to $(1 + o(1))p^{e(H)}N^{v(H)}$. The full description is now summarized in Figure 2, although we will restate it in more formal terms later on.

To see why this is a natural pseudorandomness hypothesis, we recall a famous result of Chung, Graham, and Wilson [5]. This result says that several seemingly different notions of pseudorandomness for dense graphs (i.e., graphs with constant edge density) are equivalent. These notions are based, for example, on measuring eigenvalues, edge discrepancy, subgraph counts, or codegree distributions. One rather striking fact is that having the expected 4-cycle count turns out to be equivalent to all of the other definitions.

For sparse graphs, these equivalences do not hold. While having the correct count for 4-cycles, which may be seen as the 2-blow-up of an edge (see Figure 3), still gives some control over the distribution of edges in the graph, this property is no longer strong enough to control the distribution of other small graphs such as triangles. This is where the pseudorandomness condition described above becomes useful, because knowing that we have approximately the correct count for $K_{2,2,2}$, the 2-blow-up of a triangle, does allow one to control the distribution of triangles.

We now sketch the idea behind the proof of the relative Roth theorem. We begin by noting that Roth's theorem can be rephrased as follows.

Theorem 3.1 (Roth)**.** *For every $\delta > 0$, every $A \subseteq \mathbb{Z}_N$ with $|A| \geq \delta N$ contains a 3-AP, provided N is sufficiently large.*

By a simple averaging argument (attributed to Varnavides [50]), this version of Roth's theorem is equivalent to the claim that A contains not just one, but many 3-APs.

Theorem 3.2 (Roth's theorem, counting version)**.** *For every $\delta > 0$, there exists $c = c(\delta) > 0$ such that every $A \subseteq \mathbb{Z}_N$ with $|A| \geq \delta N$ contains at least cN^2 3-APs, provided N is sufficiently large.*

To prove the relative Roth theorem from Roth's theorem, assume that $A \subseteq S \subseteq \mathbb{Z}_N$ is such that $|A| \geq \delta |S|$. The first step is to show that there is a *dense model* $\widetilde{A}$ for A. This is a dense subset of $\mathbb{Z}_N$ such that $|\widetilde{A}|/N \approx |A|/|S| \geq \delta$ and $\widetilde{A}$ approximates A in the sense of a certain cut norm. The second step is to use this cut norm condition to prove a counting lemma, which says that $\widetilde{A}$ and A contain roughly the same number of 3-APs (after an appropriate normalization), i.e.,

$$(N/|S|)^3 |\{\text{3-APs in } A\}| \approx |\{\text{3-APs in } \widetilde{A}\}|.$$

Since the counting version of Roth's theorem implies that $|\{\text{3-APs in } \widetilde{A}\}| \geq cN^2$, the relative Roth theorem is proved.

	Sets	Functions
Dense setting	$A \subseteq \mathbb{Z}_N$ $\lvert A\rvert \geq \delta N$	$f\colon \mathbb{Z}_N \to [0,1]$ $\mathbb{E}f \geq \delta$
Sparse setting	$A \subseteq S \subseteq \mathbb{Z}_N$ $\lvert A\rvert \geq \delta \lvert S\rvert$	$f \leq \nu\colon \mathbb{Z}_N \to [0,\infty)$ $\mathbb{E}f \geq \delta,\ \mathbb{E}\nu = 1+o(1)$

TABLE 1. Comparing the set version with the weighted version.

This discussion is fairly accurate, except for one white lie, which is that it is more correct to think of $\widetilde{A}$ as a weighted function from $\mathbb{Z}_N$ to $[0,1]$ than as a subset of $\mathbb{Z}_N$. It will therefore be more convenient to work with the following weighted version of Roth's theorem. At this point, it is worth fixing some notation. We will write $\mathbb{E}_{x_1\in X_1,\dots,x_k\in X_k}$ as a shorthand for $|X_1|^{-1}\dots|X_k|^{-1}\sum_{x_1\in X_1}\cdots\sum_{x_k\in X_k}$. If the variables $x_1,\dots,x_k$ or the sets $X_1,\dots,X_k$ are understood, we will sometimes choose to omit them. We will also write $o(1)$ to indicate a function that tends to zero as N tends to infinity, indicating further dependencies by subscripts when they are not understood.

Theorem 3.3 (Roth's theorem, weighted version)**.** *For every $\delta > 0$, there exists $c = c(\delta) > 0$ such that every $f\colon \mathbb{Z}_N \to [0,1]$ with $\mathbb{E}f \geq \delta$ satisfies*

$$\mathbb{E}_{x,d\in\mathbb{Z}_N}[f(x)f(x+d)f(x+2d)] \geq c - o_\delta(1). \tag{1}$$

Note that when f is $\{0,1\}$-valued, i.e., $f = 1_A$ is the indicator function of some set A, this reduces to the counting version of Roth's theorem. Up to a change of parameters, the counting version also implies the weighted version. Indeed, to deduce the weighted version from the counting version, let $A = \{x \in \mathbb{Z}_N \mid f(x) \geq \delta/2\}$. If $\mathbb{E}f \geq \delta$ and $0 \leq f \leq 1$, then $|A| \geq \delta N/2$, so

$$\mathbb{E}_{x,d\in\mathbb{Z}_N}[f(x)f(x+d)f(x+2d)] \geq (\delta/2)^3\mathbb{E}_{x,d\in\mathbb{Z}_N}[1_A(x)1_A(x+d)1_A(x+2d)].$$

By the counting version of Roth's theorem, this is bounded below by a positive constant when N is sufficiently large.

When working in the functional setting, we also replace the set S by a function $\nu\colon \mathbb{Z}_N \to [0,\infty)$. This function ν, which we call a *majorizing measure*, will be normalized to satisfy[4]

$$\mathbb{E}\nu = 1 + o(1).$$

The subset $A \subseteq S$ will be replaced by some function $f\colon \mathbb{Z}_N \to [0,\infty)$ majorized by ν, that is, such that $0 \leq f(x) \leq \nu(x)$ for all $x \in \mathbb{Z}_N$ (we write this as $0 \leq f \leq \nu$). The hypothesis $|A| \geq \delta|S|$ will be replaced by $\mathbb{E}f \geq \delta$. Note that ν and f can be unbounded, which is a major source of difficulty. The main motivating example to bear in mind is that when $A \subseteq S \subseteq \mathbb{Z}_N$, we take $\nu(x) = \frac{N}{|S|}1_S(x)$ and $f(x) = \nu(x)1_A(x)$, noting that if $|A| \geq \delta|S|$ then $\mathbb{E}f \geq \delta$. We refer the reader to Table 1 for a summary of this correspondence.

We can now state the pseudorandomness condition in a more formal way. We modify the graph G_S to a weighted graph G_ν, which, for brevity, we usually denote by ν. This is a weighted tripartite graph with vertex sets $X = Y = Z = \mathbb{Z}_N$ and edge weights given by:

- $\nu_{XY}(x,y) = \nu(2x+y)$ for all $(x,y) \in X \times Y$;
- $\nu_{XZ}(x,z) = \nu(x-z)$ for all $(x,z) \in X \times Z$;
- $\nu_{YZ}(y,z) = \nu(-y-2z)$ for all $(y,z) \in Y \times Z$.

[4]We think of ν as a sequence of functions $\nu^{(N)}$, though we usually suppress the implicit dependence of ν on N.

We will omit the subscripts if there is no risk of confusion. The pseudorandomness condition then says that the weighted graph ν has asymptotically the expected H-density for any subgraph H of $K_{2,2,2}$. For example, triangle density in ν is given by the expression $\mathbb{E}[\nu(x,y)\nu(x,z)\nu(y,z)]$, where x, y, z vary independently and uniformly over X, Y, Z, respectively. The pseudorandomness assumption requires, amongst other things, that this triangle density be $1+o(1)$, the normalization having accounted for the other factors. The full hypothesis, involving $K_{2,2,2}$ and its subgraphs, is stated below.

Definition 3.4 (3-linear forms condition)**.** A weighted tripartite graph ν with vertex sets X, Y, and Z satisfies the 3-*linear forms condition* if

$$\mathbb{E}_{x,x'\in X,\ y,y'\in Y,\ z,z'\in Z}[\nu(y,z)\nu(y',z)\nu(y,z')\nu(y',z')\nu(x,z)\nu(x',z)\nu(x,z')\nu(x',z') \cdot \nu(x,y)\nu(x',y)\nu(x,y')\nu(x',y')] = 1+o(1) \quad (2)$$

and also (2) holds when one or more of the twelve ν factors in the expectation are erased.

Similarly, a function $\nu\colon \mathbb{Z}_N \to [0,\infty)$ satisfies the 3-*linear forms condition*[5] if

$$\mathbb{E}_{x,x',y,y',z,z'\in\mathbb{Z}_N}[\nu(-y-2z)\nu(-y'-2z)\nu(-y-2z')\nu(-y'-2z')\nu(x-z)\nu(x'-z) \cdot \nu(x-z')\nu(x'-z')\nu(2x+y)\nu(2x'+y)\nu(2x+y')\nu(2x'+y')] = 1+o(1) \quad (3)$$

and also (3) holds when one or more of the twelve ν factors in the expectation are erased.

Remark. The 3-linear forms condition (3) for a function $\nu : \mathbb{Z}_N \to [0,\infty)$ is precisely the same as (2) for the weighted graph G_ν.

We can now state the relative Roth theorem formally.

Theorem 3.5 (Relative Roth)**.** *Suppose $\nu\colon \mathbb{Z}_N \to [0,\infty)$ satisfies the 3-linear forms condition. For every $\delta > 0$, there exists $c = c(\delta) > 0$ such that every $f\colon \mathbb{Z}_N \to [0,\infty)$ with $0 \le f \le \nu$ and $\mathbb{E}f \ge \delta$ satisfies*

$$\mathbb{E}_{x,d\in\mathbb{Z}_N}[f(x)f(x+d)f(x+2d)] \ge c - o_\delta(1).$$

Moreover, $c(\delta)$ may be taken to be the same constant which appears in (1)*.*

Remark. The rate at which the $o(1)$ term in (3.5) goes to zero depends not only on δ but also on the rate of convergence in the 3-linear forms condition.

4. Relative Szemerédi theorem

As in the case of Roth's theorem, we first state an equivalent version of Szemerédi's theorem allowing weights.

Theorem 4.1 (Szemerédi's theorem, weighted version)**.** *For every $k \ge 3$ and $\delta > 0$, there exists $c = c(k,\delta) > 0$ such that every $f\colon \mathbb{Z}_N \to [0,1]$ with $\mathbb{E}f \ge \delta$ satisfies*

$$\mathbb{E}_{x,d\in\mathbb{Z}_N}[f(x)f(x+d)f(x+2d)\cdots f(x+(k-1)d)] \ge c - o_{k,\delta}(1). \quad (4)$$

The setup for the relative Szemerédi theorem is a natural extension of the previous section. Just as our pseudorandomness condition for 3-APs was related to the graph-theoretic approach to Roth's theorem, the pseudorandomness condition in the general case is informed by the hypergraph removal approach to Szemerédi's theorem [20, 31, 34, 35, 46].

Instead of constructing a weighted graph as we did for 3-APs, we now construct a weighted $(k-1)$-uniform hypergraph corresponding to k-APs. For example, for 4-APs, the 3-uniform hypergraph corresponding to the majorizing measure $\nu\colon \mathbb{Z}_N \to [0,\infty)$ is 4-partite, with vertex sets W, X, Y, Z, each with N vertices labeled by elements of $\mathbb{Z}_N$. The weighted edges are given by:

[5]We will assume that N is odd, which simplifies the proof of Theorem 3.5 without too much loss in generality. Theorem 3.5 holds more generally without this additional assumption on N.

- $\nu_{WXY}(w,x,y) = \nu(3w+2x+y)$ on $W \times X \times Y$;
- $\nu_{WXZ}(w,x,z) = \nu(2w+x-z)$ on $W \times X \times Z$;
- $\nu_{WYZ}(w,y,z) = \nu(w-y-2z)$ on $W \times Y \times Z$;
- $\nu_{XYZ}(x,y,z) = \nu(-x-2y-3z)$ on $X \times Y \times Z$.

The linear forms $3w+2x+y, 2w+x-z, w-y-2z, -x-2y-3z$ are chosen because they form a 4-AP with common difference $-w-x-y-z$ and each linear form depends on exactly three of the four variables. The pseudorandomness condition then says that the weighted hypergraph ν contains asymptotically the expected count of H whenever H is a subgraph of the 2-blow-up of the simplex $K_4^{(3)}$. Here $K_4^{(3)}$ is the complete 3-uniform hypergraph on 4 vertices, that is, with vertices $\{w,x,y,z\}$ and edges $\{wxy, wxz, wyz, xyz\}$, while the 2-blow-up of $K_4^{(3)}$ is the 3-uniform hypergraph constructed by duplicating each vertex in $K_4^{(3)}$ and joining all those triples which correspond to edges in $K_4^{(3)}$. Explicitly, this 2-blow-up has vertex set $\{w_1, w_2, x_1, x_2, y_1, y_2, z_1, z_2\}$ and edges $w_i x_j y_k, w_i x_j z_k, w_i y_j z_k, x_i y_j z_k$ for all $i,j,k \in \{1,2\}$.

For general k, we are concerned with $K_k^{(k-1)}$, the complete $(k-1)$-uniform hypergraph on k vertices, while the pseudorandomness condition again asks that a certain weighted k-partite $(k-1)$-uniform hypergraph contains asymptotically the expected count for every subgraph of the 2-blow-up of $K_k^{(k-1)}$. This 2-blow-up is constructed analogously to the 2-blow-up of $K_4^{(3)}$ above and has $2k$ vertices and $k2^{k-1}$ edges.

For k-APs, the corresponding linear forms are given by the expressions $\sum_{i=1}^k (j-i)x_i$, for each $j = k, k-1, \dots, 1$. The condition (5) below is now the natural extension of the 3-linear forms condition (3). When viewed as a hypergraph condition, it asks that the count for any subgraph of the 2-blow-up of $K_k^{(k-1)}$ be close to the expected count.

Definition 4.2 (Linear forms condition)**.** A function $\nu : \mathbb{Z}_N \to [0,\infty)$ satisfies the *k-linear forms condition*[6] if

$$\mathbb{E}_{x_1^{(0)},x_1^{(1)},\dots,x_k^{(0)},x_k^{(1)} \in \mathbb{Z}_N} \Big[\prod_{j=1}^k \prod_{\omega \in \{0,1\}^{[k]\setminus\{j\}}} \nu\Big(\sum_{i=1}^k (j-i) x_i^{(\omega_i)} \Big)^{n_{j,\omega}} \Big] = 1 + o(1) \tag{5}$$

for any choice of exponents $n_{j,\omega} \in \{0,1\}$.

Now we are ready to state the main result in the proof of the Green-Tao theorem.

Theorem 4.3 (Relative Szemerédi)**.** *Suppose $k \geq 3$ and $\nu \colon \mathbb{Z}_N \to [0,\infty)$ satisfies the k-linear forms condition. For every $\delta > 0$, there exists $c = c(k,\delta) > 0$ such that every $f \colon \mathbb{Z}_N \to [0,\infty)$ with $0 \leq f \leq \nu$ and $\mathbb{E} f \geq \delta$ satisfies*

$$\mathbb{E}_{x,d \in \mathbb{Z}_N} [f(x) f(x+d) f(x+2d) \cdots f(x+(k-1)d)] \geq c - o_{k,\delta}(1). \tag{6}$$

Moreover, $c(k,\delta)$ may be taken to be the same constant which appears in (4)*.*

Remark. The rate at which the $o(1)$ term in (6) goes to zero depends not only on k and δ but also on the rate of convergence in the k-linear forms condition for ν.

Now we outline the proof of the relative Szemerédi theorem. This is simply a rephrasing of the outline given after Theorem 3.2 for the unweighted version of the relative Roth theorem. We start with $0 \leq f \leq \nu$ and $\mathbb{E} f \geq \delta$. In Section 5, we prove a *dense model theorem* which shows that there exists another function $\tilde{f} \colon \mathbb{Z}_N \to [0,1]$ which approximates f with respect to a certain cut norm.[7]

[6] As in the footnote to Definition 3.4, in our proof of Theorem 4.3 we will make the simplifying assumption that N is coprime to $(k-1)!$. In the proof of the Green-Tao theorem, one can always make this assumption.

[7] In the original Green-Tao approach, they required $\tilde{f}$ and f to be close in a stronger sense related to the Gowers uniformity norm. The cut norm approach we present here requires less stringent pseudorandomness hypotheses for applying the dense model theorem but a stronger counting lemma.

Note that $\tilde{f}$ is bounded (hence "dense" model) and $\mathbb{E}\tilde{f} \geq \delta - o(1)$. In Section 6, we establish a *counting lemma* which says that the weighted k-AP counts in f and $\tilde{f}$ are similar, that is,

$$\mathbb{E}_{x,d}[f(x)f(x+d)\cdots f(x+(k-1)d)] = \mathbb{E}_{x,d}[\tilde{f}(x)\tilde{f}(x+d)\cdots\tilde{f}(x+(k-1)d)] - o(1).$$

The right-hand side is at least $c(k,\delta) - o_{k,\delta}(1)$ by Szemerédi's theorem (Theorem 4.1). Thus the relative Szemerédi theorem follows. We now begin the proof proper.

5. Dense model theorem

Given $g\colon X \times Y \to \mathbb{R}$, viewed as an edge-weighted bipartite graph with vertex set $X \cup Y$, the *cut norm* of g, introduced by Frieze and Kannan [12] (also see [30, Chapter 8]), is defined as

$$\|g\|_\square := \sup_{A\subseteq X, B\subseteq Y} \left|\mathbb{E}_{x\in X, y\in Y}[g(x,y)1_A(x)1_B(y)]\right|. \tag{7}$$

For a weighted 3-uniform hypergraph $g\colon X \times Y \times Z \to \mathbb{R}$, we define

$$\|g\|_\square := \sup_{A\subseteq Y\times Z,\ B\subseteq X\times Z,\ C\subseteq X\times Y} \left|\mathbb{E}_{x\in X, y\in Y, z\in Z}[g(x,y,z)1_A(y,z)1_B(x,z)1_C(x,y)]\right|.$$

(The more obvious alternative, where we range A, B, C over subsets of X, Y, Z, respectively, gives a weaker norm that is not sufficient to guarantee a counting lemma.) More generally, given a weighted r-uniform hypergraph $g\colon X_1 \times \cdots \times X_r \to \mathbb{R}$, define

$$\|g\|_\square := \sup \left|\mathbb{E}_{x_1\in X_1,\ldots,x_r\in X_r}[g(x_1,\ldots,x_r)1_{A_1}(x_{-1})1_{A_2}(x_{-2})\cdots 1_{A_r}(x_{-r})]\right|, \tag{8}$$

where the supremum is taken over all choices of subsets $A_i \subseteq X_{-i} := \prod_{j\in[r]\setminus\{i\}} X_j$, $i \in [r]$, and we write

$$x_{-i} := (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_r) \in X_{-i}$$

for each i. We extend this definition of cut norm to $\mathbb{Z}_N$: for any function $f\colon \mathbb{Z}_N \to \mathbb{R}$, define

$$\|f\|_{\square,r} := \sup \left|\mathbb{E}_{x_1,\ldots,x_r\in\mathbb{Z}_N}[f(x_1+\cdots+x_r)1_{A_1}(x_{-1})1_{A_2}(x_{-2})\cdots 1_{A_r}(x_{-r})]\right|, \tag{9}$$

where the supremum is taken over all $A_1, \ldots, A_r \subseteq \mathbb{Z}_N^{r-1}$. It is easy to see that this is a norm. Equivalently, it is the hypergraph cut norm applied to the weighted r-uniform hypergraph $g\colon \mathbb{Z}_N^r \to \mathbb{R}$ with $g(x_1,\ldots,x_r) = f(x_1+\cdots+x_r)$. For example,

$$\|f\|_{\square,2} := \sup_{A,B\subseteq\mathbb{Z}_N} \left|\mathbb{E}_{x,y\in\mathbb{Z}_N}[f(x+y)1_A(x)1_B(y)]\right|.$$

The main result of this section is the following dense model theorem (in this particular form due to the third author [51]). It gives a condition under which it is possible to approximate an unbounded (or sparse) function f by a bounded (or dense) function $\tilde{f}$.

Theorem 5.1 (Dense model)**.** *For every $\epsilon > 0$, there exists an $\epsilon' > 0$ such that the following holds. Suppose $\nu\colon \mathbb{Z}_N \to [0,\infty)$ satisfies $\|\nu - 1\|_{\square,r} \leq \epsilon'$. Then, for every $f\colon \mathbb{Z}_N \to [0,\infty)$ with $f \leq \nu$, there exists a function $\tilde{f}\colon \mathbb{Z}_N \to [0,1]$ such that $\|f - \tilde{f}\|_{\square,r} \leq \epsilon$.*

Remark. One may take $\epsilon' = \exp(-\epsilon^{-C})$ where C is some absolute constant (independent of r and, more importantly, N).

A more involved dense model theorem (using a norm based on the Gowers uniformity norm rather than the cut norm) was used by Green and Tao in [23]. Its proof was subsequently simplified by Gowers [21] and, independently, Reingold, Trevisan, Tulsiani, and Vadhan [32]. Here we follow Gowers' approach, but specialized to $\|\cdot\|_{\square,r}$, which simplifies the exposition.

It will be useful to rewrite $\mathbb{E}_{x,y}[f(x+y)1_A(x)1_B(y)]$ in the form $\langle f, \varphi\rangle = \mathbb{E}_x[f(x)\varphi(x)]$ for some $\varphi\colon \mathbb{Z}_N \to \mathbb{R}$. We have, by a change of variable,

$$\mathbb{E}_{x,y}[f(x+y)1_A(x)1_B(y)] = \mathbb{E}_{x,z}[f(z)1_A(x)1_B(z-x)] = \langle f, 1_A * 1_B\rangle,$$

where the convolution is defined by $h_1 * h_2(z) := \mathbb{E}_x[h_1(x)h_2(z-x)]$. Let Φ_2 denote the set of all functions that can be written as a convex combination of convolutions $1_A * 1_B$ with $A, B \subseteq \mathbb{Z}_N$. We then have, by convexity,

$$\|f\|_{\square,2} = \sup_{A,B\subseteq\mathbb{Z}_N} |\langle f, 1_A * 1_B\rangle| = \sup_{\varphi\in\Phi_2} |\langle f,\varphi\rangle|.$$

More generally, given r functions $h_1,\dots,h_r\colon \mathbb{Z}_N^{r-1} \to \mathbb{R}$, define their generalized convolution $(h_1,\dots,h_r)^*\colon \mathbb{Z}_N \to \mathbb{R}$ by

$$(h_1,\dots,h_r)^*(x) = \mathbb{E}_{\substack{y_1,\dots,y_r\in\mathbb{Z}_N\\ y_1+\cdots+y_r=x}}[h_1(y_2,\cdots,y_r)h_2(y_1,y_3,\dots,y_r)\cdots h_r(y_1,\cdots,y_{r-1})].$$

For example, when $r=2$, we recover the usual convolution $(h_1,h_2)^* = h_1 * h_2$. We similarly have

$$\|f\|_{\square,r} = \sup_{A_1,\dots,A_r\subseteq\mathbb{Z}_N^{r-1}} |\langle f, (1_{A_1},\dots,1_{A_r})^*\rangle| = \sup_{\varphi\in\Phi_r} |\langle f,\varphi\rangle|,$$

where Φ_r is the set of all functions $\varphi : \mathbb{Z}_N \to \mathbb{R}$ that can be written as a convex combination of generalized convolutions $(1_{A_1}, 1_{A_2},\dots,1_{A_r})^*$ with $A_1,\dots,A_r \subseteq \mathbb{Z}_N^{r-1}$. The next lemma establishes a key property of Φ_r.

Lemma 5.2. *The set Φ_r is closed under multiplication, i.e., if $\varphi,\varphi' \in \Phi_r$, then $\varphi\varphi' \in \Phi_r$.*

Proof. It suffices to show that if $\varphi = (1_{A_1},\cdots,1_{A_r})^*$ and $\varphi' = (1_{B_1},\dots,1_{B_r})^*$, where $A_1,\dots,A_r$, $B_1,\dots,B_r \subseteq \mathbb{Z}_N^{r-1}$, then $\varphi\varphi' \in \Phi_r$. For any $y = (y_1,\dots,y_r) \in \mathbb{Z}_N^r$, we write $\Sigma y = y_1+\cdots+y_r$ and $y_{-i} = (y_1,\dots,y_{i-1},y_{i+1},\dots,y_r) \in \mathbb{Z}_N^{r-1}$. Then, for any $x\in\mathbb{Z}_N$, we have

$$\begin{aligned}
\varphi(x)\varphi'(x) &= \mathbb{E}_{\substack{y,y'\in\mathbb{Z}_N^r\\ \Sigma y=\Sigma y'=x}}[1_{A_1}(y_{-1})1_{B_1}(y'_{-1})\cdots 1_{A_r}(y_{-r})1_{B_r}(y'_{-r})]\\
&= \mathbb{E}_{\substack{y,z\in\mathbb{Z}_N^r\\ \Sigma y=x,\Sigma z=0}}[1_{A_1}(y_{-1})1_{B_1}(y_{-1}+z_{-1})\cdots 1_{A_r}(y_{-r})1_{B_r}(y_{-r}+z_{-r})]\\
&= \mathbb{E}_{\substack{y,z\in\mathbb{Z}_N^r\\ \Sigma y=x,\Sigma z=0}}[1_{A_1\cap(B_1-z_{-1})}(y_{-1})\cdots 1_{A_r\cap(B_r-z_{-r})}(y_{-r})]\\
&= \mathbb{E}_{\substack{z\in\mathbb{Z}_N^r\\ \Sigma z=0}}[(1_{A_1\cap(B_1-z_{-1})},\dots,1_{A_r\cap(B_r-z_{-r})})^*(x)].
\end{aligned}$$

This expresses $\varphi\varphi'$ as a convex combination of generalized convolutions. Thus $\varphi\varphi' \in \Phi_r$. □

For the rest of this section, we fix the value of r and simply write $\|\cdot\|$ for $\|\cdot\|_{\square,r}$ and Φ for Φ_r. We have $\|f\| = \sup_{\varphi\in\Phi}|\langle f,\varphi\rangle|$. An important role in the proof is played by the dual norm, which is defined by $\|\psi\|^* = \sup_{\|f\|\le 1}\langle f,\psi\rangle$. It follows easily from the definition that $|\langle f,\psi\rangle| \le \|f\|\,\|\psi\|^*$.

It is also easy to show that the unit ball for this dual norm is the convex hull of the union of Φ and $-\Phi$. To see that the convex hull is contained in the unit ball, we note that each element of $\Phi\cup(-\Phi)$ is in the unit ball and apply the triangle inequality to deduce that the same holds for convex combinations. For the reverse implication, suppose that ψ is in the unit ball of $\|\cdot\|^*$ but not in the convex hull of $\Phi\cup(-\Phi)$. Then, by the separating hyperplane theorem, there exists f such that $|\langle f,\varphi\rangle| \le 1$ for all $\varphi\in\Phi\cup(-\Phi)$ and $\langle f,\psi\rangle > 1$. But the first inequality implies that $\|f\| \le 1$ and so, by the second inequality, $\|\psi\|^* > 1$, contradicting our assumption. By Lemma 5.2, this now implies that the unit ball for the dual norm is closed under multiplication. Thus, for every $\varphi,\psi\colon \mathbb{Z}_N \to \mathbb{R}$, we have $\|(\varphi/\|\varphi\|^*)(\psi/\|\psi\|^*)\|^* \le 1$, i.e.,

$$\|\varphi\psi\|^* \le \|\varphi\|^*\,\|\psi\|^*. \tag{10}$$

Finally, we note that $\|\cdot\| \le \|\cdot\|_1$ and $\|\cdot\|_\infty \le \|\cdot\|^*$. The first inequality follows since

$$\|f\| = \sup_{\varphi\in\Phi}|\langle f,\varphi\rangle| = \sup_{\varphi\in\Phi}|\mathbb{E}_x f(x)\varphi(x)| \le \mathbb{E}_x|f(x)| = \|f\|_1.$$

The second inequality follows from duality or by letting x' be a value for which ψ achieves its maximum and taking $f(x) = N$ for $x = x'$ and 0 otherwise. It is then straightforward to verify that this function satisfies $\|f\| \le 1$ and

$$\|\psi\|^* \ge |\langle f, \psi\rangle| = |\psi(x')| = \|\psi\|_\infty .$$

Proof of Theorem 5.1. We may assume without loss of generality that $\epsilon \le \frac{1}{10}$. It suffices to show that there exists a function $\tilde{f}\colon \mathbb{Z}_N \to [0, 1+\epsilon/2]$ with $\|f - \tilde{f}\| \le \epsilon/2$. Suppose, for contradiction, that no such $\tilde{f}$ exists. Let

$$K_1 := \{\tilde{f}\colon \mathbb{Z}_N \to [0, 1+\epsilon/2]\} \quad \text{and} \quad K_2 := \{h\colon \mathbb{Z}_N \to \mathbb{R} \mid \|h\| \le \epsilon/2\}.$$

We can view K_1 and K_2 as closed convex sets in $\mathbb{R}^N$. By assumption, $f \notin K_1 + K_2 := \{\tilde{f} + h : \tilde{f} \in K_1, h \in K_2\}$. Therefore, since $K_1 + K_2$ is convex, the separating hyperplane theorem implies that there exists some $\psi\colon \mathbb{Z}_N \to \mathbb{R}$ such that

(a) $\langle f, \psi\rangle > 1$, and
(b) $\langle g, \psi\rangle \le 1$ for all $g \in K_1 + K_2$.

Note that since $0 \in K_1, K_2$, we have $K_1, K_2 \subset K_1 + K_2$. Therefore, in (b), we may take $g = (1+\epsilon/2)1_{\psi>0} \in K_1$, obtaining $\langle 1, \psi_+\rangle \le (1+\epsilon/2)^{-1}$. Here $x_+ := \max\{0, x\}$ and $\psi_+(x) := \psi(x)_+$. On the other hand, ranging g over K_2, we obtain $\|\psi\|_\infty \le \|\psi\|^* \le 2/\epsilon$, since if $\langle g, \psi\rangle \le 1$ for all g with $\|g\| \le \epsilon/2$, then $\langle g, \psi\rangle \le 2/\epsilon$ for all g with $\|g\| \le 1$.

By the Weierstrass polynomial approximation theorem, there exists some polynomial P such that $|P(x) - x_+| \le \epsilon/8$ for all $x \in [-2/\epsilon, 2/\epsilon]$. Let $P(x) = p_d x^d + \cdots + p_1 x + p_0$ and $R = |p_d|(2/\epsilon)^d + \cdots + |p_1|(2/\epsilon) + |p_0|$ (it is possible to take P so that $R = \exp(\epsilon^{-O(1)})$).

We write $P\psi$ to mean the function on $\mathbb{Z}_N$ defined by $P\psi(x) = P(\psi(x))$. Using the triangle inequality, (10), and $\|\psi\|^* \le 2/\epsilon$, we have

$$\|P\psi\|^* \le \sum_{i=0}^{d} |p_i| \|\psi^i\|^* \le \sum_{i=0}^{d} |p_i| (\|\psi\|^*)^i \le \sum_{i=0}^{d} |p_i| (2/\epsilon)^i = R.$$

Therefore, since we are assuming that $\|\nu - 1\| \le \epsilon'$,

$$|\langle \nu - 1, P\psi\rangle| \le \|\nu - 1\| \|P\psi\|^* \le \epsilon' R.$$

Since $\|\psi\|_\infty \le 2/\epsilon$, we have $\|P\psi - \psi_+\|_\infty \le \epsilon/8$. Hence,

$$\langle \nu, P\psi\rangle \le \langle 1, P\psi\rangle + \epsilon' R \le \langle 1, \psi_+\rangle + \epsilon/8 + \epsilon' R \le (1+\epsilon/2)^{-1} + \epsilon/8 + \epsilon' R.$$

Also, we have $\|\nu\|_1 = \langle \nu, 1\rangle \le \|\nu - 1\| + 1 \le 1 + \epsilon'$, where we used $\langle \nu - 1, 1\rangle \le \|\nu - 1\| \|1\|^*$ and $\|1\|^* = 1$. Thus,

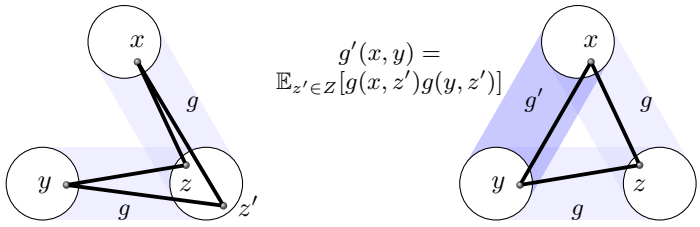
$$\langle f, \psi\rangle \le \langle f, \psi_+\rangle \le \langle \nu, \psi_+\rangle \le \langle \nu, P\psi\rangle + \|\nu\|_1 \|P\psi - \psi_+\|_\infty \le (1+\epsilon/2)^{-1} + \epsilon/8 + \epsilon' R + (1+\epsilon')\epsilon/8.$$

Since $\epsilon \le \frac{1}{10}$, the right-hand side is at most 1 when ϵ' is made sufficiently small (e.g., $\epsilon' = \epsilon/(8R)$), but this contradicts (a) from earlier. The dense model theorem follows. □

6. Counting lemma

In this section, we prove the counting lemma. We will focus principally on the graph case, Theorem 6.2 below, since this case contains all the important ideas and is notationally simpler. The hypergraph generalization is then discussed towards the end of the section.

For graphs, the counting lemma says that if two weighted graphs are close in cut norm, then they have similar triangle densities. To be more specific, we consider weighted tripartite graphs on the vertex set $X \cup Y \cup Z$, where X, Y, and Z are finite sets. Such a weighted graph g is given by three functions $g_{XY}\colon X \times Y \to \mathbb{R}$, $g_{XZ}\colon X \times Z \to \mathbb{R}$, and $g_{YZ}\colon Y \times Z \to \mathbb{R}$, although we often drop the subscripts if they are clear from context. We write $\|g\|_\square = \max\{\|g_{XY}\|_\square, \|g_{XZ}\|_\square, \|g_{YZ}\|_\square\}$.



$$\mathbb{E}[g(x,z)g(x,z')g(y,z)g(y,z')] = \mathbb{E}[g'(x,y)g(x,z)g(y,z)]$$

FIGURE 4. The densification step in the proof of the relative triangle counting lemma.

We first consider the easier case of counting in dense (i.e., bounded weight) graphs (see, for example, [30]).

Proposition 6.1 (Triangle counting lemma, dense setting)**.** *Let g and $\tilde{g}$ be weighted tripartite graphs on $X \cup Y \cup Z$ with weights in $[0,1]$. If $\|g - \tilde{g}\|_\square \le \epsilon$, then*

$$|\mathbb{E}_{x\in X, y\in Y, z\in Z}[g(x,y)g(x,z)g(y,z) - \tilde{g}(x,y)\tilde{g}(x,z)\tilde{g}(y,z)]| \le 3\epsilon.$$

Proof. Unless indicated otherwise, all expectations are taken over $x \in X$, $y \in Y$, $z \in Z$ uniformly and independently. From the definition (7) of the cut norm, we have that

$$|\mathbb{E}_{x\in X, y\in Y}[(g(x,y) - \tilde{g}(x,y))a(x)b(y)]| \le \epsilon \tag{11}$$

for every function $a\colon X \to [0,1]$ and $b\colon Y \to [0,1]$ (since the expectation is bilinear in a and b, the extrema occur when a and b are $\{0,1\}$-valued, so (11) is equivalent to (7)). It follows that

$$|\mathbb{E}[g(x,y)g(x,z)g(y,z) - \tilde{g}(x,y)g(x,z)g(y,z)]| \le \epsilon,$$

since the expectation has the form (11) if we fix any value of z. Similarly, we have

$$|\mathbb{E}[\tilde{g}(x,y)g(x,z)g(y,z) - \tilde{g}(x,y)\tilde{g}(x,z)g(y,z)]| \le \epsilon$$

and

$$|\mathbb{E}[\tilde{g}(x,y)\tilde{g}(x,z)g(y,z) - \tilde{g}(x,y)\tilde{g}(x,z)\tilde{g}(y,z)]| \le \epsilon.$$

The result then follows from telescoping and the triangle inequality. □

This proof does not work in the sparse setting, when g is unbounded, since (11) requires a and b to be bounded. The main result of this section, stated next for graphs (the hypergraph version is stated towards the end of the section), gives a counting lemma assuming $0 \le g \le \nu$ for some ν satisfying the linear forms condition. This is one of the main results in our paper [7].

Theorem 6.2 (Relative triangle counting lemma)**.** *Let $\nu, g, \tilde{g}$ be weighted tripartite graphs on $X \cup Y \cup Z$. Assume that ν satisfies the 3-linear forms condition (Definition 3.4), $0 \le g \le \nu$, and $0 \le \tilde{g} \le 1$. If $\|g - \tilde{g}\|_\square = o(1)$, then*

$$|\mathbb{E}_{x\in X, y\in Y, z\in Z}[g(x,y)g(x,z)g(y,z) - \tilde{g}(x,y)\tilde{g}(x,z)\tilde{g}(y,z)]| = o(1).$$

The proof uses repeated application of the Cauchy-Schwarz inequality, a standard technique in this area, popularized by Gowers [17, 18, 19, 20]. The key additional idea, introduced in [7, 8], is *densification* (see Figure 4). After several applications of the Cauchy-Schwarz inequality, it becomes necessary to analyze the 4-cycle density: $\mathbb{E}_{x,y,z,z'}[g(x,z)g(x,z')g(y,z)g(y,z')]$. To do this, one introduces an auxiliary weighted graph $g'\colon X \times Y \to [0,\infty)$ defined by $g'(x,y) := \mathbb{E}_{z'}[g(x,z')g(y,z')]$ (this is basically the codegree function). Note that we benefit here from working with weighted graphs. The expression for the 4-cycle density now becomes $\mathbb{E}_{x,y,z}[g'(x,y)g(x,z)g(y,z)]$.

At first glance, it seems that our reasoning is circular. Our aim was to estimate a certain triangle density expression but we have now returned to another triangle density expression. However, g' behaves much more like a dense weighted graph with bounded edge weights, so what we have

accomplished is to replace one of the "sparse" g_{XY} by a "dense" g'_{XY}. If we do this two more times, replacing g_{YZ} and g_{XZ} with dense counterparts, the problem reduces to the dense case, which we already know how to handle.

We begin with a warm-up showing how to apply the Cauchy-Schwarz inequality (there will be many more applications later on). The following lemma shows that the 3-linear forms condition on ν implies $\|\nu - 1\|_\square = o(1)$, which we need to apply the dense model theorem, Theorem 5.1.

Lemma 6.3. *For any $\nu\colon X \times Y \to \mathbb{R}$,*

$$\|\nu - 1\|_\square \le (\mathbb{E}_{x,x' \in X, y,y' \in Y}[(\nu(x,y)-1)(\nu(x',y)-1)(\nu(x,y')-1)(\nu(x',y')-1)])^{1/4}. \tag{12}$$

Remark. The right-hand side of (12) is the Gowers uniformity norm of $\nu - 1$. The lemma shows that the cut norm is weaker than the Gowers uniformity norm. To see $\|\nu - 1\|_\square = o(1)$, we expand the right-hand side of (12) into an alternating sum of linear forms in ν, each being $1 + o(1)$ by the linear forms condition, so that the alternating sum cancels to $o(1)$.

Proof. By repeated applications of the Cauchy-Schwarz inequality, we have, for $A \subseteq X$ and $B \subseteq Y$,

$$\begin{aligned}
|\mathbb{E}_{x,y}[(\nu(x,y)-1)1_A(x)1_B(y)]|^4 &\le \left|\mathbb{E}_x[(\mathbb{E}_y[(\nu(x,y)-1)1_B(y)])^2 1_A(x)]\right|^2 \\
&\le \left|\mathbb{E}_x[(\mathbb{E}_y[(\nu(x,y)-1)1_B(y)])^2]\right|^2 \\
&= \left|\mathbb{E}_{x,y,y'}[(\nu(x,y)-1)(\nu(x,y')-1)1_B(y)1_B(y')]\right|^2 \\
&\le \mathbb{E}_{y,y'}[(\mathbb{E}_x[(\nu(x,y)-1)(\nu(x,y')-1)])^2 1_B(y)1_B(y')] \\
&\le \mathbb{E}_{y,y'}[(\mathbb{E}_x[(\nu(x,y)-1)(\nu(x,y')-1)])^2] \\
&= \mathbb{E}_{x,x',y,y'}[(\nu(x,y)-1)(\nu(x',y)-1)(\nu(x,y')-1)(\nu(x',y')-1)].
\end{aligned}$$

The lemma then follows. □

The next lemma is crucial to what follows. It shows that in certain expressions a factor ν can be deleted from an expectation while incurring only a $o(1)$ loss.

Lemma 6.4 (Strong linear forms)**.** *Let $\nu, g, \tilde{g}$ be weighted tripartite graphs on $X \cup Y \cup Z$. Assume that ν satisfies the 3-linear forms condition, $0 \le g \le \nu$, and $0 \le \tilde{g} \le 1$. Then*

$$\mathbb{E}_{x \in X, y \in Y, z,z' \in Z}[(\nu(x,y)-1)g(x,z)g(x,z')g(y,z)g(y,z')] = o(1)$$

and the same statement holds if any subset of the four g factors are replaced by $\tilde{g}$.

Proof. We give the proof when none of the g factors are replaced. The other cases require only a simple modification. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left|\mathbb{E}_{x,y,z,z'}[(\nu(x,y)-1)g(x,z)g(x,z')g(y,z)g(y,z')]\right|^2 \\
&\le \mathbb{E}_{y,z,z'}[(\mathbb{E}_x[(\nu(x,y)-1)g(x,z)g(x,z')])^2 g(y,z)g(y,z')]\ \mathbb{E}_{y,z,z'}[g(y,z)g(y,z')] \\
&\le \mathbb{E}_{y,z,z'}[(\mathbb{E}_x[(\nu(x,y)-1)g(x,z)g(x,z')])^2 \nu(y,z)\nu(y,z')]\ \mathbb{E}_{y,z,z'}[\nu(y,z)\nu(y,z')].
\end{aligned}$$

The second factor is at most $1 + o(1)$ by the linear forms condition. So it remains to analyze the first factor. We have, by another application of the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\mathbb{E}_{y,z,z'}[(\mathbb{E}_x[(\nu(x,y)-1)g(x,z)g(x,z')])^2\nu(y,z)\nu(y,z')]\right|^2 \\
&= \left|\mathbb{E}_{x,x',y,z,z'}[(\nu(x,y)-1)(\nu(x',y)-1)g(x,z)g(x,z')g(x',z)g(x',z')\nu(y,z)\nu(y,z')]\right|^2 \\
&= \left|\mathbb{E}_{x,x',z,z'}[\mathbb{E}_y[(\nu(x,y)-1)(\nu(x',y)-1)\nu(y,z)\nu(y,z')]g(x,z)g(x,z')g(x',z)g(x',z')]\right|^2 \\
&\le \mathbb{E}_{x,x',z,z'}[(\mathbb{E}_y[(\nu(x,y)-1)(\nu(x',y)-1)\nu(y,z)\nu(y,z')])^2 g(x,z)g(x,z')g(x',z)g(x',z')] \\
&\qquad \cdot \mathbb{E}_{x,x',z,z'}[g(x,z)g(x,z')g(x',z)g(x',z')] \\
&\le \mathbb{E}_{x,x',z,z'}[(\mathbb{E}_y[(\nu(x,y)-1)(\nu(x',y)-1)\nu(y,z)\nu(y,z')])^2 \nu(x,z)\nu(x,z')\nu(x',z)\nu(x',z')] \\
&\qquad \cdot \mathbb{E}_{x,x',z,z'}[\nu(x,z)\nu(x,z')\nu(x',z)\nu(x',z')].
\end{aligned}$$

Using the 3-linear forms condition, the second factor is $1+o(1)$ and the first factor is $o(1)$ (expand everything and observe that all the terms are $1 + o(1)$ and the signs make all the 1's cancel). □

Proof of Theorem 6.2. If ν is identically 1, we are in the dense setting, in which case the theorem follows from Proposition 6.1. Now we apply induction on the number of $\nu_{XY}, \nu_{XZ}, \nu_{YZ}$ which are identically 1. By relabeling if necessary, we may assume without loss of generality that ν_{XY} is not identically 1. We define auxiliary weighted graphs $\nu', g', \tilde{g}' \colon X \times Y \to [0,\infty)$ by

$$\begin{aligned}
\nu'(x,y) &:= \mathbb{E}_z[\nu(x,z)\nu(y,z)], \\
g'(x,y) &:= \mathbb{E}_z[g(x,z)g(y,z)], \\
\tilde{g}'(x,y) &:= \mathbb{E}_z[\tilde{g}(x,z)\tilde{g}(y,z)].
\end{aligned}$$

We refer to this step as *densification*. The idea is that even though ν and g are possibly unbounded, the new weighted graphs ν' and g' behave like dense graphs. The weights on ν' and g' are not necessarily bounded by 1, but they almost are. We cap the weights by setting $g'_{\wedge 1} := \min\{g', 1\}$ and $\nu'_{\wedge 1} := \min\{\nu', 1\}$ and show that the capping has negligible effect. We have

$$\mathbb{E}[g(x,y)g(x,z)g(y,z) - \tilde{g}(x,y)\tilde{g}(x,z)\tilde{g}(y,z)] = \mathbb{E}[gg' - \tilde{g}\tilde{g}'] = \mathbb{E}[g(g'-\tilde{g}')] + \mathbb{E}[(g-\tilde{g})\tilde{g}'], \quad (13)$$

where the first expectation is taken over $x \in X, y \in Y, z \in Z$ and the other expectations are taken over $X \times Y$ (we will use these conventions unless otherwise specified). The second term on the right-hand side of (13) equals $\mathbb{E}[(g(x,y) - \tilde{g}(x,y))\tilde{g}(x,z)\tilde{g}(y,z)]$ and its absolute value is at most $\|g - \tilde{g}\|_\square = o(1)$ (here we use $0 \le \tilde{g} \le 1$ as in the proof of Proposition 6.1). So it remains to bound the first term on the right-hand side of (13). By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
(\mathbb{E}[g(g'-\tilde{g}')])^2 \le \mathbb{E}[g(g'-\tilde{g}')^2]\ \mathbb{E}[g] &\le \mathbb{E}[\nu(g'-\tilde{g}')^2]\ \mathbb{E}[\nu] \\
&= \mathbb{E}_{x,y}[\nu(x,y)(\mathbb{E}_z[g(x,z)g(y,z) - \tilde{g}(x,z)\tilde{g}(y,z)])^2]\ \mathbb{E}_{x,y}[\nu(x,y)].
\end{aligned}$$

The second factor is $1 + o(1)$ by the linear forms condition. By Lemma 6.4, the first factor differs from

$$\mathbb{E}_{x,y}[(\mathbb{E}_z[g(x,z)g(y,z) - \tilde{g}(x,z)\tilde{g}(y,z)])^2] = \mathbb{E}[(g'-\tilde{g}')^2] \quad (14)$$

by $o(1)$ (take the difference, expand the square, and then apply Lemma 6.4 term-by-term).

The 3-linear forms condition implies that $\mathbb{E}[\nu'] = 1 + o(1)$ and $\mathbb{E}[\nu'^2] = 1 + o(1)$. Therefore, by the Cauchy-Schwarz inequality, we have

$$(\mathbb{E}[|\nu' - 1|])^2 \le \mathbb{E}[(\nu'-1)^2] = o(1). \quad (15)$$

We want to show that (14) is $o(1)$. We have

$$\mathbb{E}[(g'-\tilde{g}')^2] = \mathbb{E}[(g'-\tilde{g}')(g'-g'_{\wedge 1})] + \mathbb{E}[(g'-\tilde{g}')(g'_{\wedge 1}-\tilde{g}')]. \quad (16)$$

Since $0 \le g' \le \nu'$, we have

$$0 \le g' - g'_{\wedge 1} = \max\{g'-1, 0\} \le \max\{\nu'-1, 0\} \le |\nu'-1|. \quad (17)$$

Using (15) and (17), the absolute value of the first term on the right-hand side of (16) is at most

$$\mathbb{E}[(\nu'+1)|\nu'-1|] = \mathbb{E}[(\nu'-1)|\nu'-1|] + 2\mathbb{E}[|\nu'-1|] = o(1).$$

Next, we claim that

$$\|g'_{\wedge 1} - \tilde{g}'\|_\square = o(1). \tag{18}$$

Indeed, for any $A \subseteq X$ and $B \subseteq Y$, we have

$$\mathbb{E}_{x,y}[(g'_{\wedge 1} - \tilde{g}')(x,y)1_A(x)1_B(y)] = \mathbb{E}[(g'_{\wedge 1} - \tilde{g}')1_{A\times B}] = \mathbb{E}[(g'_{\wedge 1} - g')1_{A\times B}] + \mathbb{E}[(g' - \tilde{g}')1_{A\times B}].$$

By (17) and (15), the absolute value of the first term is at most $\mathbb{E}[|\nu' - 1|] = o(1)$. The second term can be rewritten as

$$\mathbb{E}_{x,y,z}[1_{A\times B}(x,y)g(x,z)g(y,z) - 1_{A\times B}(x,y)\tilde{g}(x,z)\tilde{g}(y,z)],$$

which is $o(1)$ by the induction hypothesis (replace $\nu_{XY}, g_{XY}, \tilde{g}_{XY}$ by 1, $1_{A\times B}$, $1_{A\times B}$, respectively, and note that this increases the number of $\{\nu_{XY}, \nu_{XZ}, \nu_{YZ}\}$ which are identically 1). This proves (18).

We now expand the second term on the right-hand side of (16) as

$$\mathbb{E}[(g' - \tilde{g}')(g'_{\wedge 1} - \tilde{g}')] = \mathbb{E}[g'g'_{\wedge 1}] - \mathbb{E}[g'\tilde{g}'] - \mathbb{E}[\tilde{g}'g'_{\wedge 1}] + \mathbb{E}[\tilde{g}'^2]. \tag{19}$$

We claim that each of the expectations on the right-hand side is $\mathbb{E}[(\tilde{g}')^2] + o(1)$. Indeed, we have

$$\mathbb{E}[g'g'_{\wedge 1}] - \mathbb{E}[(\tilde{g}')^2] = \mathbb{E}_{x,y,z}[g'_{\wedge 1}(x,y)g(x,z)g(y,z) - \tilde{g}'(x,y)\tilde{g}(x,z)\tilde{g}(y,z)],$$

which is $o(1)$ by the induction hypothesis (replace $\nu_{XY}, g_{XY}, \tilde{g}_{XY}$ by $1, g'_{\wedge 1}, \tilde{g}'$, respectively, which by (18) satisfies $\|g'_{\wedge 1} - \tilde{g}'\|_\square = o(1)$, and note that this increases the number of $\{\nu_{XY}, \nu_{XZ}, \nu_{YZ}\}$ which are identically 1). One can similarly show that the other expectations on the right-hand side of (19) are also $\mathbb{E}[(\tilde{g}')^2] + o(1)$. Thus (19) is $o(1)$ and the theorem follows. □

The main difficulty in extending Theorem 6.2 to hypergraphs is notational. As discussed in Section 4, to study k-APs, we consider $(k-1)$-uniform k-partite weighted hypergraphs. The vertex sets will be denoted $X_1, \dots, X_k$ (in application $X_i = \mathbb{Z}_N$ for all i). We write $X_{-i} := X_1 \times \cdots \times X_{i-1} \times X_{i+1} \times \cdots \times X_k$ and $x_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k)$ for any $x = (x_1, \dots, x_k) \in X_1 \times \cdots \times X_k$. Then a weighted hypergraph g consists of functions $g_{-i} \colon X_{-i} \to \mathbb{R}$ for each $i = 1, \dots, k$. As before, we drop the subscripts if they are clear from context. We write $\|g\|_\square = \max\{\|g_{-1}\|_\square, \dots, \|g_{-k}\|_\square\}$, where $\|g_{-i}\|_\square$ is the cut norm of g_{-i} defined in (8).

The appropriate generalization of the 3-linear forms condition involves counts for the 2-blow-up of the simplex $K_k^{(k-1)}$. We say that a weighted hypergraph ν satisfies the *k-linear forms condition* (the hypergraph version of Definition 4.2) if

$$\mathbb{E}_{x_1^{(0)}, x_1^{(1)} \in X_1, \dots, x_k^{(0)}, x_k^{(1)} \in X_k} \Big[\prod_{j=1}^{k} \prod_{\omega \in \{0,1\}^{[k]\setminus\{j\}}} \nu(x_{-j}^{(\omega)}) \Big] = 1 + o(1)$$

and also the same statement holds if any subset of the ν factors (there are $k2^{k-1}$ such factors) are deleted. Here $x_{-j}^{(\omega)} := (x_1^{(\omega_1)}, \dots, x_{j-1}^{(\omega_{j-1})}, x_{j+1}^{(\omega_{j+1})}, \dots, x_k^{(\omega_k)}) \in X_{-j}$.

The following theorem generalizes Theorem 6.2.

Theorem 6.5 (Relative simplex counting lemma)**.** *Let $\nu, g, \tilde{g}$ be weighted $(k-1)$-uniform k-partite weighted hypergraphs on $X_1 \cup \cdots \cup X_k$. Assume that ν satisfies the k-linear forms condition, $0 \le g \le \nu$ and $0 \le \tilde{g} \le 1$. If $\|g - \tilde{g}\|_\square = o(1)$, then*

$$|\mathbb{E}_{x_1 \in X_1, \dots, x_k \in X_k}[g(x_{-1})g(x_{-2})\cdots g(x_{-k}) - \tilde{g}(x_{-1})\tilde{g}(x_{-2})\cdots\tilde{g}(x_{-k})]| = o(1).$$

The proof of Theorem 6.5 is a straightforward generalization of the proof of Theorem 6.2. We simply point out the necessary modifications and leave the reader to figure out the details (a full proof can be found in our paper [7], but it is perhaps easier to reread the graph case and think about the small changes that need to be made).

The proof proceeds by induction on the number of $\nu_{-1}, \dots, \nu_{-k}$ which are not identically 1. When $\nu = 1$, we are in the dense setting and the proof of Proposition 6.1 easily extends. Now assume that ν_{-1} is not identically 1. We have extensions of Lemmas 6.3 and 6.4, where in the proof we have to apply the Cauchy-Schwarz inequality $k-1$ times in succession. For the densification step, we define $\nu', g', \tilde{g}' \colon X_{-1} \to [0,\infty)$ by

$$\begin{aligned}\nu'(x_{-1}) &= \mathbb{E}_{x_1 \in X_1}[\nu(x_{-2}) \cdots \nu(x_{-k})],\\ g'(x_{-1}) &= \mathbb{E}_{x_1 \in X_1}[g(x_{-2}) \cdots g(x_{-k})],\\ \tilde{g}'(x_{-1}) &= \mathbb{E}_{x_1 \in X_1}[\tilde{g}(x_{-2}) \cdots \tilde{g}(x_{-k})].\end{aligned}$$

The rest of the proof works with minimal changes.

7. Proof of the relative Szemerédi theorem

We are now ready to prove the relative Szemerédi theorem using the dense model theorem and the counting lemma following the outline given in Section 4.

Proof of Theorem 4.3. The k-linear forms condition implies that $\|\nu - 1\|_{\square,k-1} = o(1)$ (by a sequence of $k-1$ applications of the Cauchy-Schwarz inequality, following Lemma 6.3). By the dense model theorem, Theorem 5.1, we can find $\tilde{f} \colon \mathbb{Z}_N \to [0,1]$ so that $\|f - \tilde{f}\|_{\square,k-1} = o(1)$.

Let $X_1 = X_2 = \cdots = X_k = \mathbb{Z}_N$. For each $j = 1, \dots, k$, define the linear form $\psi_j \colon X_{-j} \to \mathbb{Z}_N$ by

$$\psi_j(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k) := \sum_{i \in [k] \setminus \{j\}} (j-i) x_i.$$

Construct $(k-1)$-uniform k-partite weighted hypergraphs $\nu, g, \tilde{g}$ on $X_1 \cup \cdots \cup X_k$ by setting

$$\nu_{-j}(x_{-j}) := \nu(\psi_j(x_{-j})), \quad g_{-j}(x_{-j}) := f(\psi_j(x_{-j})), \quad \tilde{g}_{-j}(x_{-j}) := \tilde{f}(\psi_j(x_{-j}))$$

(in the first definition, the left ν_{-j} refers to the weighted hypergraph and the second ν refers to the given function on $\mathbb{Z}_N$). We claim that

$$\|\nu_{-j} - 1\|_{\square} = \|\nu - 1\|_{\square,k-1} \tag{20}$$

and

$$\|g_{-j} - \tilde{g}_{-j}\|_{\square} = \|f - \tilde{f}\|_{\square,k-1} \tag{21}$$

for every j (in both (20) and (21) the left-hand side refers to the hypergraph cut norm (8) while the right-hand side refers to the cut norm (9) for functions on $\mathbb{Z}_N$). We illustrate (21) in the case when $k = j = 4$ (the full proof is straightforward). The left-hand side of (21) equals

$$\sup_{A_1,A_2,A_3 \subseteq \mathbb{Z}_N^2} \left| \mathbb{E}_{x_1,x_2,x_3 \in \mathbb{Z}_N}[(f - \tilde{f})(3x_1 + 2x_2 + x_3) 1_{A_1}(x_2,x_3) 1_{A_2}(x_1,x_3) 1_{A_3}(x_1,x_2)] \right|, \tag{22}$$

while the right-hand side of (21) equals

$$\sup_{B_1,B_2,B_3 \subseteq \mathbb{Z}_N^2} \left| \mathbb{E}_{x_1,x_2,x_3 \in \mathbb{Z}_N}[(f - \tilde{f})(x_1 + x_2 + x_3) 1_{B_1}(x_2,x_3) 1_{B_2}(x_1,x_3) 1_{B_3}(x_1,x_2)] \right|. \tag{23}$$

These two expressions are equal[8] up to a change of variables $3x_1 \leftrightarrow x_1$ and $2x_2 \leftrightarrow x_2$.

[8]Here we use the assumption in the footnote to Definition 4.2 that N is coprime to $(k-1)!$. Without this assumption, it can be shown that the two norms (22) and (23) differ by at most a constant factor depending on k, which would also suffice for what follows.

It follows from (21) that $\|g-\tilde{g}\|_\square = \|f-\tilde{f}\|_{\square,k-1} = o(1)$. Moreover, the k-linear forms condition for $\nu\colon \mathbb{Z}_N \to [0,\infty)$ translates to the k-linear forms condition for the weighted hypergraph ν. It follows from the counting lemma, Theorem 6.5, that

$$\mathbb{E}_{x_1,\dots,x_k\in\mathbb{Z}_N^k}[g_{-1}(x_{-1})\cdots g_{-k}(x_{-k})] = \mathbb{E}_{x_1,\dots,x_k\in\mathbb{Z}_N^k}[\tilde{g}_{-1}(x_{-1})\cdots \tilde{g}_{-k}(x_{-k})] + o(1). \tag{24}$$

The left-hand side is equal to

$$\mathbb{E}_{x_1,\dots,x_k\in\mathbb{Z}_N^k}[f(\psi_1(x_{-1}))\cdots f(\psi_k(x_{-k}))] = \mathbb{E}_{x,d\in\mathbb{Z}_N}[f(x)f(x+d)\cdots f(x+(k-1)d)],$$

which can be seen by setting $x=\psi_1(x_{-1})$ and $d = x_1+\cdots+x_k$ so that $\psi_j(x_{-j}) = x+(j-1)d$. A similar statement holds for the right-hand side of (24). So (24) is equivalent to

$$\mathbb{E}_{x,d\in\mathbb{Z}_N}[f(x)f(x+d)\cdots f(x+(k-1)d)] = \mathbb{E}_{x,d\in\mathbb{Z}_N}[\tilde{f}(x)\tilde{f}(x+d)\cdots \tilde{f}(x+(k-1)d)] + o(1),$$

which is at least $c(k,\delta) - o_{k,\delta}(1)$ by Theorem 4.1, as desired. □

8. Constructing the majorant

In this section, we use the relative Szemerédi theorem to prove the Green-Tao theorem. To do this, we must construct a majorizing measure for the primes that satisfies the linear forms condition.

Rather than considering the set of primes itself, we put weights on the primes, a common technique in analytic number theory. The weights we use will be related to the well-known von Mangoldt function Λ. This is defined by $\Lambda(n) = \log p$ if $n = p^k$ for some prime p and positive integer k and $\Lambda(n) = 0$ if n is not a power of a prime (actually, the higher powers $p^2, p^3, \dots$ play no role here and we will soon discard them from Λ). That these are natural weights to consider follows from the observation that the Prime Number Theorem is equivalent to $\sum_{n\le N}\Lambda(n) = (1+o(1))N$.

A difficulty with using Λ is that it is biased on certain residue classes. For example, every prime other than 2 is odd. This prevents us from making any pseudorandomness claims unless we can somehow remove these biases. This is achieved using the *W-trick*. Let $w = w(N)$ be any function that tends to infinity slowly with N. Let $W = \prod_{p\le w} p$ be the product of primes up to w. The trick for avoiding biases mod p for any $p \le w$ is to consider only those primes which are congruent to 1 (mod W). In keeping with this idea, we define the modified von Mangoldt function by

$$\widetilde{\Lambda}(n) := \begin{cases} \frac{\phi(W)}{W}\log(Wn+1) & \text{when } Wn+1 \text{ is prime,} \\ 0 & \text{otherwise.} \end{cases}$$

The factor $\phi(W)/W$ is present since exactly $\phi(W)$ of the W residue classes mod W have infinitely many primes and a strong form of Dirichlet's theorem[9] tells us that the primes are equidistributed among these $\phi(W)$ residue classes, i.e., $\sum_{n\le N}\widetilde{\Lambda}(n) = (1+o(1))N$ as long as w grows slowly enough with N. From now on, we will work with $\widetilde{\Lambda}$ rather than Λ. Our main goal is to prove the following result, which says that there is a majorizing measure for $\widetilde{\Lambda}$ which satisfies the linear forms condition.

Proposition 8.1. *For every $k \ge 3$, there exists $\delta_k > 0$ such that for every sufficiently large N there exists a function $\nu\colon \mathbb{Z}_N \to [0,\infty)$ satisfying the k-linear forms condition and $\nu(n) \ge \delta_k\widetilde{\Lambda}(n)$ for all $N/2 \le n < N$.*

[9]In fact, Dirichlet's theorem, or even the Prime Number Theorem, are not necessary to prove the Green-Tao theorem, though we assume them to simplify the exposition. Indeed, a weaker form of the Prime Number Theorem asserting that there are at least $cN/\log N$ primes up to N for some $c > 0$ suffices for our needs (this bound was first proved by Chebyshev and a famous short proof was subsequently found by Erdős; see [1, Ch. 2]). Furthermore, in place of Dirichlet's theorem, a simple pigeonhole argument shows that for each W, some residue class b (mod W) contains many primes (whereas we use Dirichlet's theorem to take $b = 1$). The proof presented here can easily be modified to deal with general b, though the notation gets a bit more cumbersome as b could vary with W. An analysis of this sort is necessary to prove a Szemerédi-type statement for the primes (see Section 10), since we do not then know how our subset of the primes is distributed on congruence classes.

Using this majorant with the relative Szemerédi theorem, we obtain the Green-Tao theorem.

Proof of Theorem 1.1 assuming Proposition 8.1. Define $f\colon \mathbb{Z}_N \to [0,\infty)$ by $f(n) = \delta_k \widetilde{\Lambda}(n)$ if $N/2 \le n < N$ and $f(n) = 0$ otherwise. By Dirichlet's theorem, $\sum_{N/2\le n<N} f(n) = (1/2 + o(1))\delta_k N$, so $\mathbb{E}f \ge \delta_k/3$ for large N. Since $0 \le f \le \nu$ and ν satisfies the k-linear forms condition, it follows from the relative Szemerédi theorem, Theorem 4.3, that $\mathbb{E}[f(x)f(x+d)\cdots f(x+(k-1)d)] \ge c(k,\delta_k/3) - o_{k,\delta}(1)$. Therefore, for sufficiently large N, we have $f(x)f(x+d)\cdots f(x+(k-1)d) > 0$ for some $N/2 \le x < N$ and $d \ne 0$ (the $d = 0$ terms contribute negligibly to the expectation). Since f is supported on $[N/2, N)$, we see that $x, x+d, \dots, x+(k-1)d$ is not only an AP in $\mathbb{Z}_N$ but also in $\mathbb{Z}$, i.e., has no wraparound issues. Thus $(x+jd)W + 1$ for $j = 0, \dots, k-1$ is a k-AP of primes. □

How do we construct the majorant ν for $\widetilde{\Lambda}(n)$? Recall that the Möbius function μ is defined by $\mu(n) = (-1)^{\omega(n)}$ when n is square-free, where $\omega(n)$ is the number of prime factors of n, and $\mu(n) = 0$ when n is not square-free. The functions Λ and μ are related by the Möbius inversion formula

$$\Lambda(n) = \sum_{d|n} \mu(d)\log(n/d).$$

In Green and Tao's original proof, the following truncated version of Λ (motivated by [16]) was used to construct the majorant. For any $R > 0$, define

$$\Lambda_R(n) := \sum_{\substack{d|n \\ d\le R}} \mu(d)\log(R/d).$$

Observe that if n has no prime divisors less than or equal to R, then $\Lambda_R(n) = \log R$. Tao [42] later simplified the proof by using the following variant of Λ_R, where the restriction $d \le R$ is replaced by a smoother cutoff.

Definition 8.2. Let $\chi\colon \mathbb{R} \to [0,1]$ be any smooth, compactly supported function. Define

$$\Lambda_{\chi,R}(n) := \log R \sum_{d|n} \mu(d)\chi\left(\frac{\log d}{\log R}\right).$$

In our application, χ will be supported on $[-1,1]$, so only divisors d which are at most R are considered in the sum. Note that Λ_R above corresponds to $\chi(x) = \max\{1-|x|, 0\}$, which is not smooth. The following proposition, which we will prove in the next section, gives a linear forms estimate for $\Lambda_{\chi,R}$.

Proposition 8.3 (Linear forms estimate)**.** *Fix any smooth function $\chi\colon \mathbb{R} \to [0,1]$ supported on $[-1,1]$. Let m and t be positive integers. Let $\psi_1, \dots, \psi_m\colon \mathbb{Z}^t \to \mathbb{Z}$ be fixed linear maps, with no two being multiples of each other. Assume that $R = o(N^{1/(10m)})$ grows with N and w grows sufficiently slowly with N. Let $W := \prod_{p\le w} p$. Write $\theta_i := W\psi_i + 1$. Let B be a product $\prod_{i=1}^t I_i$, where each I_i is a set of at least R^{10m} consecutive integers. Then*

$$\mathbb{E}_{\mathbf{x}\in B}[\Lambda_{\chi,R}(\theta_1(\mathbf{x}))^2 \cdots \Lambda_{\chi,R}(\theta_m(\mathbf{x}))^2] = (1+o(1))\left(\frac{Wc_\chi \log R}{\phi(W)}\right)^m, \tag{25}$$

where $o(1)$ denotes a quantity tending to zero as $N \to \infty$ (at a rate that may depend on χ, m, t, $\psi_1, \dots, \psi_m$, R, and w), and c_χ is the normalizing factor

$$c_\chi := \int_0^\infty |\chi'(x)|^2\, dx.$$

Now we construct the majorizing measure ν and show that it satisfies the linear forms condition.

Proposition 8.4. *Fix any smooth function $\chi\colon \mathbb{R} \to [0,1]$ supported on $[-1,1]$ with $\chi(0) = 1$. Let $k \geq 3$ and $R := N^{k^{-1}2^{-k-3}}$. Assume that w grows sufficiently slowly with N and let $W := \prod_{p \leq w} p$. Define $\nu\colon \mathbb{Z}_N \to [0,\infty)$ by*

$$\nu(n) := \begin{cases} \frac{\phi(W)}{W} \frac{\Lambda_{\chi,R}(Wn+1)^2}{c_\chi \log R} & \text{when } N/2 \leq n < N, \\ 1 & \text{otherwise.} \end{cases} \tag{26}$$

Then ν satisfies the k-linear forms condition.

Note that while $\Lambda_{\chi,R}$ is not necessarily nonnegative, ν constructed in (26) is always nonnegative due to the square on $\Lambda_{\chi,R}$.

Proof of Proposition 8.1 assuming Proposition 8.4. Take $\delta_k = k^{-1}2^{-k-4}c_\chi^{-1}$. It suffices to verify that for N sufficiently large we have $\delta_k \widetilde{\Lambda}(n) \leq \nu(n)$ for all $N/2 \leq n < N$. We only need to check the inequality when $Wn+1$ is prime, since $\widetilde{\Lambda}(n)$ is zero otherwise. We have

$$\log R = k^{-1}2^{-k-3} \log N \geq k^{-1}2^{-k-4} \log(WN+1) = c_\chi \delta_k \log(WN+1),$$

where the inequality holds for sufficiently large N provided w grows slowly enough. When $Wn+1$ is prime, we have $\Lambda_{\chi,R}(Wn+1) = \log R$, so

$$\delta_k \widetilde{\Lambda}(n) = \delta_k \frac{\phi(W)}{W} \log(Wn+1) \leq \delta_k \frac{\phi(W)}{W} \log(WN+1) \leq \frac{\phi(W)}{W} \frac{\log R}{c_\chi} = \nu(n),$$

as claimed. □

Proof of Proposition 8.4 assuming Proposition 8.3. We need to check that

$$\mathbb{E}_{\mathbf{x} \in \mathbb{Z}_N^t} [\nu(\psi_1(\mathbf{x})) \cdots \nu(\psi_m(\mathbf{x}))] = 1 + o(1) \tag{27}$$

whenever $\psi_1, \ldots, \psi_m$, $m \leq k2^{k-1}$, are the linear forms that appear in (5) or any subset thereof. Note that no two ψ_i are multiples of each other.

To use the two-piece definition of ν, we divide the domain $\mathbb{Z}_N$ into intervals. Let $Q = Q(N)$ be a slowly increasing function of N. Divide $\mathbb{Z}_N$ into Q roughly equal intervals and form a partition of $\mathbb{Z}_N^t$ into Q^t boxes, as follows:

$$B_{u_1,\ldots,u_t} = \prod_{j=1}^{t} ([u_j N/Q, (u_j+1)N/Q) \cap \mathbb{Z}_N) \subseteq \mathbb{Z}_N^t, \qquad u_1, \ldots, u_t \in \mathbb{Z}_Q.$$

Then, up to a $o(1)$ error (due to the fact that the boxes do not all have exactly equal sizes), the left-hand side of (27) equals

$$\mathbb{E}_{u_1,\ldots,u_t \in \mathbb{Z}_Q} [\mathbb{E}_{\mathbf{x} \in B_{u_1,\ldots,u_t}} [\nu(\psi_1(\mathbf{x})) \cdots \nu(\psi_m(\mathbf{x}))]].$$

We say that a box $B_{u_1,\ldots,u_t}$ is *good* if, for each $j \in [m]$, the set $\{\psi_j(\mathbf{x}) : \mathbf{x} \in B_{u_1,\ldots,u_t}\}$ either lies completely in the subset $[N/2, N)$ of $\mathbb{Z}_N$ or completely outside this subset. Otherwise, we say that the box is *bad*. We may assume Q grows slowly enough that $N/Q \geq R^{10m}$. From Proposition 8.3 and the definition of ν, we know that for good boxes,

$$\mathbb{E}_{\mathbf{x} \in B_{u_1,\ldots,u_t}} [\nu(\psi_1(\mathbf{x})) \cdots \nu(\psi_m(\mathbf{x}))] = 1 + o(1).$$

For bad boxes, we use the bound $\nu(n) \leq 1 + \frac{\phi(W)}{W} \frac{\Lambda_{\chi,R}(Wn+1)^2}{c_\chi \log R}$. By expanding and applying (25) to each term, we find that

$$\mathbb{E}_{\mathbf{x} \in B_{u_1,\ldots,u_t}} [\nu(\psi_1(\mathbf{x})) \cdots \nu(\psi_m(\mathbf{x}))] = O(1)$$

(it is bounded in absolute value by $2^m + o(1)$). It remains to show that the proportion of boxes that are bad is $o(1)$.

Suppose $B_{u_1,\dots,u_t}$ is bad. Then there exists some i such that the image of the box under ψ_i intersects both $[N/2, N)$ and its complement. This implies that there exists some (real-valued) $\mathbf{x} \in \prod_{j=1}^t [u_j N/Q, (u_j+1)N/Q) \subseteq (\mathbb{R}/N\mathbb{Z})^t$ with $\psi_i(\mathbf{x}) = 0$ or $N/2 \pmod N$. Letting $\mathbf{y} = Q\mathbf{x}/N$, we see that $\mathbf{y} \in \prod_{j=1}^t [u_j, u_j+1) \subseteq (\mathbb{R}/Q\mathbb{Z})^t$ satisfies $\psi_i(\mathbf{y}) = 0$ or $Q/2 \pmod Q$. This implies that $\psi_i(u_1, \dots, u_t)$ is either $O(1)$ or $Q/2 + O(1) \pmod Q$. Since ψ_i is a nonzero linear form, at most a $O(1/Q)$ fraction of the tuples $(u_1, \dots, u_t) \in \mathbb{Z}_Q^t$ have this property. This can be seen by noting that if we fix all but one of the coordinates, there will be $O(1)$ choices for the final coordinate for which ψ_i is in the required range. Taking the union over all i, we see that the proportion of bad boxes is $O(1/Q) = o(1)$. □

9. Verifying the linear forms condition

In this section, we prove Proposition 8.3. There are numerous estimates along the way. To avoid getting bogged down with the rather technical error bounds, we first go through the proof while skipping some of these details (i.e., by only considering the "main term"). The approximations are then justified at the end, where we collect the error bound arguments. We note that all constants will depend implicitly on $\chi, m, t, \psi_1, \dots, \psi_m$.

Expanding the definition of $\Lambda_{\chi,R}$, we rewrite the left-hand side of (25) as

$$(\log R)^{2m} \sum_{d_1,d_1',\dots,d_m,d_m' \in \mathbb{N}} \left(\prod_{j=1}^m \mu(d_j)\chi\left(\frac{\log d_j}{\log R}\right) \mu(d_j')\chi\left(\frac{\log d_j'}{\log R}\right) \right) \mathbb{E}_{\mathbf{x}\in B}[1_{d_j,d_j' | \theta_j(\mathbf{x}) \ \forall j}]. \tag{28}$$

Since $\mu(d) = 0$ unless d is square-free, we only need to consider square-free $d_1, d_1', \dots, d_m, d_m'$. Also, since χ is supported on $[-1, 1]$, we may assume that $d_1, d_1', \dots, d_m, d_m' \le R$. Let D denote the lcm of $d_1, d_1', \dots, d_m, d_m'$. The width of the box B is at least R^{10m} in each dimension, so, by considering a slightly smaller box $B' \subseteq B$ such that each dimension of B' is divisible by $D \le R^{2m}$, we obtain

$$\mathbb{E}_{\mathbf{x}\in B}[1_{d_j,d_j' | \theta_j(\mathbf{x}) \, \forall j}] = \mathbb{E}_{\mathbf{x}\in \mathbb{Z}_D^t}[1_{d_j,d_j' | \theta_j(\mathbf{x}) \, \forall j}] + O(R^{-8m}).$$

Therefore, as there are at most R^{2m} choices for $d_1, d_1', \dots, d_m, d_m'$, we see that, up to an additive error of $O(R^{-6m} \log^{2m} R)$, we may approximate (28) by

$$(\log R)^{2m} \sum_{d_1,d_1',\dots,d_m,d_m' \in \mathbb{N}} \left(\prod_{j=1}^m \mu(d_j)\chi\left(\frac{\log d_j}{\log R}\right) \mu(d_j')\chi\left(\frac{\log d_j'}{\log R}\right) \right) \mathbb{E}_{\mathbf{x}\in \mathbb{Z}_D^t}[1_{d_j,d_j' | \theta_j(\mathbf{x}) \ \forall j}]. \tag{29}$$

Let φ be the Fourier transform of $e^x\chi(x)$. That is,

$$e^x \chi(x) = \int_{\mathbb{R}} \varphi(\xi) e^{-ix\xi} \, d\xi.$$

Substituting and simplifying, we have

$$\chi\left(\frac{\log d}{\log R}\right) = \int_{\mathbb{R}} d^{-\frac{1+i\xi}{\log R}} \varphi(\xi) \, d\xi.$$

We wish to plug this integral into (29). It helps to first restrict the integral to a compact interval $I = [-\log^{1/2} R, \log^{1/2} R]$. By basic results in Fourier analysis (see, for example, [40, Chapter 5, Theorem 1.3]), since χ is smooth and compactly supported, φ decays rapidly, that is, $\varphi(\xi) = O_A((1+|\xi|)^{-A})$ for any $A > 0$. It follows that for any $A > 0$,

$$\chi\left(\frac{\log d}{\log R}\right) = \int_I d^{-\frac{1+i\xi}{\log R}} \varphi(\xi) \, d\xi + O_A(d^{-1/\log R} (\log R)^{-A}). \tag{30}$$

We write

$$z_j := \frac{1+i\xi_j}{\log R} \qquad \text{and} \qquad z_j' := \frac{1+i\xi_j'}{\log R}.$$

We have $\chi(\log d/\log R) = O(d^{-1/\log R})$ (we only need to check this for $d \le R$ since χ is supported on $[-1,1]$). Using (30), we have

$$\prod_{j=1}^m \chi\left(\frac{\log d_j}{\log R}\right)\chi\left(\frac{\log d_j'}{\log R}\right) = \int_I \cdots \int_I \prod_{j=1}^m d_j^{-z_j} {d_j'}^{-z_j'} \varphi(\xi_j)\varphi(\xi_j')\, d\xi_j d\xi_j' + O_A\left((\log R)^{-A}\prod_{j=1}^m (d_j d_j')^{-1/\log R}\right). \quad (31)$$

Using (31), we estimate (29) (error bounds are deferred to the end) by

$$(\log R)^{2m} \int_I \cdots \int_I \sum_{d_1,d_1',\ldots,d_m,d_m' \in \mathbb{N}} \mathbb{E}_{\mathbf{x}\in\mathbb{Z}_D^t}[1_{d_j,d_j'|\theta_j(\mathbf{x})\ \forall j}] \prod_{j=1}^m \mu(d_j) d_j^{-z_j} \mu(d_j') {d_j'}^{-z_j'} \varphi(\xi_j)\varphi(\xi_j')\, d\xi_j d\xi_j'. \quad (32)$$

We are allowed to swap the summation and the integrals because I is compact and the sum can be shown to be absolutely convergent (the argument for absolute convergence is similar to the error bound for (32) included towards the end of the section). Splitting $d_1, d_1', \ldots, d_m, d_m'$ in (32) into prime factors, we obtain

$$(32) = (\log R)^{2m} \int_I \cdots \int_I \prod_p E_p(\xi) \cdot \prod_{j=1}^m \varphi(\xi_j)\varphi(\xi_j')\, d\xi_j d\xi_j', \quad (33)$$

where $\xi = (\xi_1, \xi_1', \ldots, \xi_m, \xi_m') \in I^{2m}$ and $E_p(\xi)$ is the Euler factor

$$E_p(\xi) := \sum_{d_1,d_1',\ldots,d_m,d_m' \in \{1,p\}} \mathbb{E}_{\mathbf{x}\in\mathbb{Z}_p^t}[1_{d_j,d_j'|\theta_j(\mathbf{x})\ \forall j}] \prod_{j=1}^m \mu(d_j) d_j^{-z_j} \mu(d_j') {d_j'}^{-z_j'}.$$

We have $E_p(\xi) = 1$ when $p \le w$ (recall the W-trick, so $p \nmid \theta_j(\mathbf{x}) = W\psi_j(\mathbf{x}) + 1$ for all j when $p \le w$). When $p > w$, the expectation in the summand equals 1 if all d_j, d_j' are 1, $1/p$ if $d_j d_j' = 1$ for all except exactly one j, and is at most $1/p^2$ otherwise (here we assume that w is sufficiently large so that no two ψ_i are multiples of each other mod p). It follows that for $p > w$,

$$E_p(\xi) = 1 - p^{-1}\sum_{j=1}^m (p^{-z_j} + p^{-z_j'} - p^{-z_j - z_j'}) + O(p^{-2}) = (1 + O(p^{-2}))E_p'(\xi),$$

where, for any prime p,

$$E_p'(\xi) := \prod_{j=1}^m \frac{(1 - p^{-1-z_j})(1 - p^{-1-z_j'})}{1 - p^{-1-z_j-z_j'}}.$$

It then follows that

$$\prod_p E_p(\xi) = \prod_{p>w}(1 + O(p^{-2}))E_p'(\xi) = (1 + O(w^{-1}))\left(\prod_{p\le w} E_p'(\xi)\right)^{-1} \prod_p E_p'(\xi). \quad (34)$$

Recall that the Riemann zeta function

$$\zeta(s) := \sum_{n\ge 1} n^{-s} = \prod_p (1 - p^{-s})^{-1}$$

has a simple pole at $s = 1$ with residue 1 (a proof is included towards the end). This implies that

$$\prod_p E_p'(\xi) = \prod_{j=1}^m \frac{\zeta(1 + z_j + z_j')}{\zeta(1 + z_j)\zeta(1 + z_j')} \approx \prod_{j=1}^m \frac{z_j z_j'}{z_j + z_j'}, \quad (35)$$

where $\approx$ denotes asymptotic equality. Here we use $|z_1|, |z_1'|, \dots, |z_m|, |z_m'| = O((\log R)^{-1/2})$ as $\xi_1, \xi_1', \dots \xi_m, \xi_m' \in I$. For $p \le w$, we make the approximation $E_p'(\xi) \approx (1-p^{-1})^m$. Hence,

$$\prod_{p \le w} E_p'(\xi) \approx \prod_{p \le w} (1-p^{-1})^m = \left(\frac{\phi(W)}{W}\right)^m. \tag{36}$$

Substituting (34), (35), and (36) into (33), we find that

$$(33) \approx (\log R)^{2m} \left(\frac{W}{\phi(W)}\right)^m \int_I \cdots \int_I \prod_{j=1}^m \frac{z_j z_j'}{z_j + z_j'} \varphi(\xi_j)\varphi(\xi_j')\, d\xi_j d\xi_j'. \tag{37}$$

It remains to estimate the integral

$$\int_I \int_I \frac{z_j z_j'}{z_j + z_j'} \varphi(\xi_j)\varphi(\xi_j')\, d\xi_j d\xi_j' = \frac{1}{\log R} \int_I \int_I \frac{(1+i\xi_j)(1+i\xi_j')}{2 + i(\xi_j + \xi_j')} \varphi(\xi_j)\varphi(\xi_j')\, d\xi_j d\xi_j'.$$

We can replace the domain of integration $I = [-\log^{1/2} R, \log^{1/2} R]$ by $\mathbb{R}$ with a loss of $O_A(\log^{-A} R)$ for any $A > 0$ due to the rapid decay of φ given by $\varphi(\xi) = O_A((1+|\xi|)^{-A})$. We claim

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(1+i\xi)(1+i\xi')}{2 + i(\xi + \xi')} \varphi(\xi)\varphi(\xi')\, d\xi d\xi' = \int_0^\infty |\chi'(x)|^2\, dx = c_\chi. \tag{38}$$

Using

$$\frac{1}{2 + i(\xi + \xi')} = \int_0^\infty e^{-(1+i\xi)x} e^{-(1+i\xi')x}\, dx,$$

we can rewrite the left-hand side of (38) as

$$\int_0^\infty \left(\int_{\mathbb{R}} \varphi(\xi)(1+i\xi) e^{-(1+i\xi)x}\, d\xi \right)^2 dx.$$

The expression in parentheses is $-\chi'(x)$, so (38) follows. Substituting (38) into (37) we arrive at the desired conclusion, Proposition 8.3.

Error estimates. Now we bound the error terms in the above analysis.

Simple pole of Riemann zeta function. Here is the argument showing that $\zeta(s) = (s-1)^{-1} + O(1)$ whenever $\operatorname{Re} s > 1$ and $s - 1 = O(1)$. We have $(s-1)^{-1} = \int_1^\infty x^{-s}\, dx$. So

$$\zeta(s) - \frac{1}{s-1} = \sum_{n=1}^\infty n^{-s} - \int_1^\infty x^{-s}\, dx = \sum_{n=1}^\infty \int_n^{n+1} (n^{-s} - x^{-s})\, dx.$$

The n-th term on the right is bounded in magnitude by $O(n^{-2})$. So the sum is $O(1)$.

Estimate (32). We want to bound the difference between (32) and (29). This means bounding the contribution to (29) from the error term in (31). Taking absolute values everywhere, we bound these contributions by

$$\begin{aligned} &\ll_A (\log R)^{O(1)-A} \sum_{\substack{d_1, d_1', \dots, d_m, d_m' \\ \text{sq-free integers}}} \mathbb{E}_{\mathbf{x} \in \mathbb{Z}_D^t} [1_{d_j, d_j' | \theta_j(\mathbf{x}) \ \forall j}] (d_1 d_1' \cdots d_m d_m')^{-1/\log R} \\ &= (\log R)^{O(1)-A} \prod_p \sum_{d_1, d_1', \dots, d_m, d_m' \in \{1, p\}} \mathbb{E}_{\mathbf{x} \in \mathbb{Z}_p^t} [1_{d_j, d_j' | \theta_j(\mathbf{x}) \ \forall j}] (d_1 d_1' \cdots d_m d_m')^{-1/\log R}. \end{aligned}$$

The expectation $\mathbb{E}_{\mathbf{x}\in\mathbb{Z}_p^t}[1_{d_j,d_j'|\theta_j(\mathbf{x})\ \forall j}]$ is 1 if all d_i and d_i' are 1 and at most $1/p$ otherwise. We continue to bound the above by

$$\begin{aligned}
&\leq (\log R)^{O(1)-A}\prod_p\left(1+p^{-1}\sum_{\substack{d_1,d_1',\dots,d_m,d_m'\in\{1,p\}\\ \text{not all 1's}}}(d_1d_1'\cdots d_md_m')^{-1/\log R}\right)\\
&= (\log R)^{O(1)-A}\prod_p\left(1+p^{-1}((p^{-1/\log R}+1)^{2m}-1)\right)\\
&\leq (\log R)^{O(1)-A}\prod_p\left(1-p^{-1-1/\log R}\right)^{-O(1)}\\
&= (\log R)^{O(1)-A}\zeta(1+1/\log R)^{O(1)}.
\end{aligned}$$

So the difference between (32) and (29) is $O_A((\log R)^{O(1)-A})$, which is small as long as we take A to be sufficiently large.

Estimate in (35). We have $|z_j|, |z_j'| = O(\log^{-1/2} R)$ since $|\xi_j|, |\xi_j'| \leq \log^{1/2} R$. So

$$\prod_{j=1}^m \frac{\zeta(1+z_j+z_j')}{\zeta(1+z_j)\zeta(1+z_j')} = \prod_{j=1}^m \frac{((z_j+z_j')^{-1}+O(1))}{(z_j^{-1}+O(1))(z_j'^{-1}+O(1))} = (1+O(\log^{-1/2} R))\prod_{j=1}^m \frac{z_jz_j'}{z_j+z_j'}. \quad (39)$$

Estimate in (36). If $|z|\log p = O(1)$ (which is the case for $p \leq w$), then

$$1-p^{-1-z} = 1-p^{-1}e^{-z\log p} = 1-p^{-1}(1+O(|z|\log p)) = (1-p^{-1})(1+O(|z|p^{-1}\log p)).$$

It follows that for all $p \leq w$ and $\xi_1, \xi_1', \dots, \xi_m, \xi_m' \in I$, we have

$$E_p'(\xi) = \left(1+O\left(\frac{\log p}{p\log^{1/2} R}\right)\right)(1-p^{-1})^m$$

and, hence,

$$\prod_{p\leq w} E_p'(\xi) = \left(1+O\left(\frac{w}{\log^{1/2} R}\right)\right)\prod_{p\leq w}(1-p^{-1})^m. \quad (40)$$

Estimate in (37). Using (34), (39), and (40), we find that the ratio between the two sides in (37) is $1+O(1/w+w/\log^{1/2} R) = 1+o(1)$, as long as w grows sufficiently slowly.

10. Extensions of the Green-Tao theorem

We conclude by discussing a few extensions of the Green-Tao theorem.

Szemerédi's theorem in the primes. As noted already by Green and Tao [23], their method also implies a Szemerédi-type theorem for the primes. That is, every subset of the primes with positive relative upper density contains arbitrarily long arithmetic progressions.

One elegant corollary of this result is that there are arbitrarily long APs where every term is a sum of two squares. This result follows from a combination of the well-known fact that every prime of the form $4n+1$ is a sum of two squares with Dirichlet's theorem on primes in arithmetic progressions, which tells us that roughly half the primes are congruent to 1 (mod 4). Even this innocent-sounding corollary was open before Green and Tao's paper.

Gaussian primes contain arbitrarily shaped constellations. The Gaussian integers is the set of all numbers of the form $a + bi$, where $a, b \in \mathbb{Z}$. This set is a ring under the usual definitions of addition and multiplication for complex numbers. It is also a unique factorization domain, so it is legitimate to talk about the set of Gaussian primes. Tao [44] proved that an analogue of the Green-Tao theorem holds for the Gaussian primes.

We say that $A \subseteq \mathbb{Z}^d$ *contains arbitrary constellations* if, for every finite set $F \subseteq \mathbb{Z}^d$, there exist $x \in \mathbb{Z}^d$ and $t \in \mathbb{Z}_{>0}$ such that $x + tf \in A$ for every $f \in F$. Tao's theorem then states that the Gaussian primes, viewed as a subset of $\mathbb{Z}^2$, contain arbitrary constellations. Just as the Green-Tao theorem uses Szemerédi's theorem as a black box, this theorem uses the multidimensional analogue of Szemerédi's theorem, first proved by Furstenberg and Katznelson [14]. This states that every subset of $\mathbb{Z}^d$ with positive upper density[10] contains arbitrary constellations. The Furstenberg-Katznelson theorem also follows from the hypergraph removal lemma and the approach taken by Tao is to transfer this hypergraph removal proof to the sparse context. It may therefore be seen as a precursor to the approach taken here.

Multidimensional Szemerédi theorem in the primes. Let P denote the set of primes in $\mathbb{Z}$. It was shown recently by Tao and Ziegler [48] and, independently, by Cook, Magyar, and Titichetrakun [9], that every subset of P^d of positive relative upper density contains arbitrary constellations. A short proof was subsequently given in [11] (though, like [48], it assumes some difficult results of Green, Tao, and Ziegler that we will discuss later in this section).

Although both this result and Tao's result on the Gaussian primes are multidimensional analogues of the Green-Tao theorem, they are quite different in nature. Informally speaking, a key difficulty in the second result is that there is a strong correlation between coordinates in P^d (namely, that all coordinates are simultaneously prime), whereas there is no significant correlation between the real and imaginary parts of a typical Gaussian prime (after applying an extension of the W-trick).

The primes contain arbitrary polynomial progressions. We say that $A \subseteq \mathbb{Z}$ *contains arbitrary polynomial progressions* if, whenever $P_1, \dots, P_k \in \mathbb{Z}[X]$ are polynomials in one variable with integer coefficients satisfying $P_1(0) = \cdots = P_k(0) = 0$, there is some $x \in \mathbb{Z}$ and $t \in \mathbb{Z}_{>0}$ such that $x + P_j(t) \in A$ for each $j = 1, \dots, k$. A striking generalization of Szemerédi's theorem due to Bergelson and Leibman [3] states that any subset of $\mathbb{Z}$ of positive upper density contains arbitrary polynomial progressions. To date, the only known proofs of this result use ergodic theory.

For primes, an analogue of the Bergelson-Leibman theorem was proved by Tao and Ziegler [49]. This result states that any subset of the primes with positive relative upper density contains arbitrary polynomial progressions. In particular, the primes themselves contain arbitrary polynomial progressions. It seems plausible that the simplifications outlined here could also be used to simplify the proof of this theorem.

The number of k-APs in the primes. The original approach of Green and Tao (and the approach outlined in this paper) implies that for any k the number of k-APs of primes with each term at most N is on the order of $\frac{N^2}{\log^k N}$. In subsequent work, Green, Tao, and Ziegler [25, 26, 27] showed how to determine the exact asymptotic. That is, they determine a constant c_k such that the number of k-APs of primes with each term at most N is $(c_k + o(1))\frac{N^2}{\log^k N}$. More generally, they determine an asymptotic for the number of prime solutions to a broad range of linear systems of equations.

The proof of these results also draws on the transference technique discussed in this paper but a number of additional ingredients are needed, most notably an inverse theorem describing the structure of those sets which do not contain the expected number of solutions to certain linear

[10] A set $A \subseteq \mathbb{Z}^d$ is said to have positive upper density if $\limsup_{N\to\infty} \left|A \cap [-N, N]^d\right| / (2N+1)^d > 0$. We say that $A \subseteq S \subseteq \mathbb{Z}^d$ has positive relative upper density if $\limsup_{N\to\infty} \left|A \cap S \cap [-N, N]^d\right| / \left|S \cap [-N, N]^d\right| > 0$.

systems of equations. It is this result which is transferred to the sparse setting when one wishes to determine the exact asymptotic.

Acknowledgments. We thank Yuval Filmus, Mohammad Bavarian, and the anonymous referee for helpful comments on the manuscript.

References

[1] M. Aigner and G. M. Ziegler, *Proofs from The Book*, fourth ed., Springer-Verlag, Berlin, 2010.
[2] F. A. Behrend, *On sets of integers which contain no three terms in arithmetical progression*, Proc. Nat. Acad. Sci. U.S.A. **32** (1946), 331–332.
[3] V. Bergelson and A. Leibman, *Polynomial extensions of van der Waerden's and Szemerédi's theorems*, J. Amer. Math. Soc. **9** (1996), 725–753.
[4] T. F. Bloom, *A quantitative improvement for Roth's theorem on arithmetic progressions*, J. Lond. Math. Soc. (2) **93** (2016), 643–663.
[5] F. R. K. Chung, R. L. Graham, and R. M. Wilson, *Quasi-random graphs*, Combinatorica **9** (1989), 345–362.
[6] D. Conlon and J. Fox, *Graph removal lemmas*, Surveys in Combinatorics 2013, London Math. Soc. Lecture Note Ser., Cambridge University Press, 2013, pp. 1–50.
[7] D. Conlon, J. Fox, and Y. Zhao, *A relative Szemerédi theorem*, Geom. Funct. Anal. **25** (2015), 733–762.
[8] D. Conlon, J. Fox, and Y. Zhao, *Extremal results in sparse pseudorandom graphs*, Adv. Math. **256** (2014), 206–290.
[9] B. Cook, A. Magyar, and T. Titichetrakun, *A multidimensional Szemerédi theorem in the primes*, arXiv:1306.3025.
[10] J. Fox, *A new proof of the graph removal lemma*, Ann. of Math. **174** (2011), 561–579.
[11] J. Fox and Y. Zhao, *A short proof of the multidimensional Szemerédi theorem in the primes*, Amer. J. Math. **137** (2015), 1139–1145.
[12] A. Frieze and R. Kannan, *Quick approximation to matrices and applications*, Combinatorica **19** (1999), 175–220.
[13] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.
[14] H. Furstenberg and Y. Katznelson, *An ergodic Szemerédi theorem for commuting transformations*, J. Analyse Math. **34** (1978), 275–291.
[15] H. Furstenberg, Y. Katznelson, and D. Ornstein, *The ergodic theoretical proof of Szemerédi's theorem*, Bull. Amer. Math. Soc. **7** (1982), 527–552.
[16] D. A. Goldston and C. Y. Yıldırım, *Higher correlations of divisor sums related to primes. I. Triple correlations*, Integers **3** (2003), A5, 66pp.
[17] W. T. Gowers, *A new proof of Szemerédi's theorem for arithmetic progressions of length four*, Geom. Funct. Anal. **8** (1998), 529–551.
[18] ———, *A new proof of Szemerédi's theorem*, Geom. Funct. Anal. **11** (2001), 465–588.
[19] ———, *Quasirandomness, counting and regularity for 3-uniform hypergraphs*, Combin. Probab. Comput. **15** (2006), 143–184.
[20] ———, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Ann. of Math. **166** (2007), 897–946.
[21] ———, *Decompositions, approximate structure, transference, and the Hahn-Banach theorem*, Bull. Lond. Math. Soc. **42** (2010), 573–606.
[22] B. Green, *Long arithmetic progressions of primes*, Analytic number theory, Clay Math. Proc., vol. 7, Amer. Math. Soc., Providence, RI, 2007, pp. 149–167.
[23] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. **167** (2008), 481–547.
[24] ———, *New bounds for Szemerédi's theorem. II. A new bound for $r_4(N)$*, Analytic number theory, Cambridge Univ. Press, Cambridge, 2009, pp. 180–204.
[25] ———, *Linear equations in primes*, Ann. of Math. **171** (2010), 1753–1850.
[26] ———, *The Möbius function is strongly orthogonal to nilsequences*, Ann. of Math. **175** (2012), 541–566.
[27] B. Green, T. Tao, and T. Ziegler, *An inverse theorem for the Gowers $U^{s+1}[N]$-norm*, Ann. of Math. **176** (2012), 1231–1372.
[28] B. Host, *Progressions arithmétiques dans les nombres premiers (d'après B. Green et T. Tao)*, Astérisque (2006), Exp. No. 944, viii, 229–246, Séminaire Bourbaki. Vol. 2004/2005.
[29] B. Kra, *The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view*, Bull. Amer. Math. Soc. **43** (2006), 3–23.
[30] L. Lovász, *Large networks and graph limits*, Amer. Math. Soc. Colloq. Publ., vol. 60, Amer. Math. Soc., Providence, RI, 2012.

[31] B. Nagle, V. Rödl, and M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms **28** (2006), 113–179.
[32] O. Reingold, L. Trevisan, M. Tulsiani, and S. Vadhan, *New proofs of the Green-Tao-Ziegler dense model theorem: an exposition*, arXiv:0806.0381.
[33] ———, *Dense subsets of pseudorandom sets*, 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, 2008, pp. 76–85.
[34] V. Rödl and J. Skokan, *Regularity lemma for k-uniform hypergraphs*, Random Structures Algorithms **25** (2004), 1–42.
[35] ———, *Applications of the regularity lemma for uniform hypergraphs*, Random Structures Algorithms **28** (2006), 180–194.
[36] K. F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 104–109.
[37] I. Z. Ruzsa and E. 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, 1978, pp. 939–945.
[38] T. Sanders, *On Roth's theorem on progressions*, Ann. of Math. **174** (2011), 619–636.
[39] T. Schoen and I. D. Shkredov, *Roth's theorem in many variables*, Israel J. Math. **199** (2014), 287–308.
[40] E. M. Stein and R. Shakarchi, *Fourier analysis: an introduction,* Princeton Lectures in Analysis, 1, Princeton University Press, Princeton, NJ, 2003.
[41] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 199–245.
[42] T. Tao, *A remark on Goldston-Yıldırım correlation estimates*, available at `http://www.math.ucla.edu/~tao/preprints/Expository/gy-corr.dvi`.
[43] ———, *Arithmetic progressions and the primes*, Collect. Math. (2006), 37–88.
[44] ———, *The Gaussian primes contain arbitrarily shaped constellations*, J. Anal. Math. **99** (2006), 109–176.
[45] ———, *Obstructions to uniformity and arithmetic patterns in the primes*, Pure Appl. Math. Q. **2** (2006), 395–433.
[46] ———, *A variant of the hypergraph removal lemma*, J. Combin. Theory Ser. A **113** (2006), 1257–1280.
[47] ———, *The dichotomy between structure and randomness, arithmetic progressions, and the primes*, International Congress of Mathematicians. Vol. I, Eur. Math. Soc., Zürich, 2007, pp. 581–608.
[48] T. Tao and T. Ziegler, *A multi-dimensional Szemerédi theorem for the primes via a correspondence principle*, Israel J. Math. **207** (2015), 203–228.
[49] ———, *The primes contain arbitrarily long polynomial progressions*, Acta Math. **201** (2008), 213–305.
[50] P. Varnavides, *On certain sets of positive density*, J. London Math. Soc. **34** (1959), 358–360.
[51] Y. Zhao, *An arithmetic transference proof of a relative Szemerédi theorem*, Math. Proc. Cambridge Philos. Soc. **156** (2014), 255–261.

Mathematical Institute, Oxford OX2 6GG, United Kingdom
E-mail address: `david.conlon@maths.ox.ac.uk`

Department of Mathematics, MIT, Cambridge, MA 02139-4307
E-mail address: `fox@math.mit.edu`

Department of Mathematics, MIT, Cambridge, MA 02139-4307
E-mail address: `yufeiz@math.mit.edu`