ARITHMETIC PROGRESSIONS AND THE PRIMES - EL ESCORIAL LECTURES

TERENCE TAO

Abstract. We describe some of the machinery behind recent progress in establishing infinitely many arithmetic progressions of length k in various sets of integers, in particular in arbitrary dense subsets of the integers, and in the primes.

1. Introduction

A celebrated theorem of Roth [36] in 1953 asserts:

Theorem 1.1 (Roth's theorem, first version)**.** [36] *Let $A \subset \mathbb{Z}^+$ be a subset of integers with positive upper density, thus* $\limsup_{N\to\infty} \frac{1}{N}|A \cap [1,N]| > 0$. *Then A contains infinitely many arithmetic progressions $n, n+r, n+2r$ of length three.*

Here we of course restrict the spacing of the progression r to be non-negative. This theorem was originally proven by Roth by Fourier analytic methods and a stopping time argment, and we shall reprove it below (in fact, we shall give two proofs). This theorem was then generalized substantially by Szemerédi in 1975:

Theorem 1.2 (Szemerédi's theorem, first version)**.** [38], [39] *Let $A \subset \mathbb{Z}^+$ be a subset of integers with positive upper density, thus* $\limsup_{N\to\infty} \frac{1}{N}|A \cap [1,N]| > 0$, *and let $k \geqslant 3$. Then A contains infinitely many arithmetic progressions $n, n+r, \ldots, n+(k-1)r$ of length k.*

Thus Roth's theorem is the $k = 3$ version of Szemerédi's theorem. (The cases $k < 3$ are trivial).

Szemerédi's original proof was combinatorial (relying in particular on graph theory) and very complicated. A substantially shorter proof - but one involving the full machinery of measure theory and ergodic theory, as well as the axiom of choice - was obtained by Furstenberg [10], [11] in 1977. Since then, there have been two other types of proofs; a proof of Gowers [16], [17] in 2001 which combines "higher order" Fourier analytic methods with techniques from additive combinatorics; and also arguments of Gowers [18] and Rodl-Skokan [34], [35] using the machinery of hypergraphs. While we will not discuss all these separate proofs in detail here, we will need to discuss certain ideas from each of these arguments as they will eventually be used in the proof of Theorem 5.1 below.

The above theorems do not apply directly to the set of prime numbers, as they have density zero. Nevertheless, in 1939 van der Corput [43] proved, by using Fourier analytic methods (the Hardy-Littlewood circle method) which were somewhat similar to the methods used by Roth, the following result:

1991 *Mathematics Subject Classification.* 11N13, 11B25, 374A5.
The author is supported by a grant from the Packard Foundation.

Theorem 1.3 (Van der Corput's theorem)**.** [43] *Let $P \subset \mathbb{Z}^+$ be the set of primes. Then P contains infinitely many arithmetic progressions $n, n+r, n+2r$ of length three.*

However, just as Roth's Fourier-analytic methods proved very difficult to extend beyond the $k = 3$ case, so too did van der Corput's arguments. The proof relied on very delicate information concerning the Fourier coefficients of the primes (or more precisely of the von Mangoldt function $\Lambda(n)$, which is essentially supported on the primes). This additional information allows one to not only show that there are infinitely many progressions of primes of length three, but also to obtain an asymptotic count as to *how many* such progressions there are; we shall return to this point later.

Roth's theorem and van der Corput's theorem were combined by Green [20] in 2003 to obtain

Theorem 1.4 (Green's theorem)**.** [20] *Let $A \subset P$ be a subset of primes with positive relative upper density:*

$$\limsup_{N \to \infty} \frac{|A \cap [1, N]|}{|P \cap [1, N]|} > 0.$$

Then A contains infinitely many arithmetic progressions $n, n+r, n+2r$ of length three.

A key observation made in that paper was that one did not need very deep number-theoretic information about the structure of A or P to prove this result. In fact, the same result holds not just for relatively dense subsets of primes, but relatively dense subsets of *almost primes* (numbers containing no small prime factors); we shall return to this point later.

In 2004, Ben Green and the author [23] were able to extend this theorem to arbitrarily long progressions, by replacing Fourier-analytic ideas with ergodic theory ones:

Theorem 1.5. [24] *Let $A \subset P$ be a subset of primes with positive relative upper density:*

$$\limsup_{N \to \infty} \frac{|A \cap [1, N]|}{|P \cap [1, N]|} > 0,$$

and let $k \geqslant 3$. Then A contains infinitely many arithmetic progressions $n, n+r, \ldots, n+(k-1)r$ of length k. In particular, the primes contain arbitrarily long arithmetic progressions.

At the time of writing, we are not able to obtain van der Corput's more precise asymptotic estimate on the *number* of prime progressions of arbitrary length k, but we are able to do so in the $k = 4$ case; see Section 6.

In this expository article, we review briefly the methods of proof of Roth's theorem and Szemerédi's theorem for various values of k, focusing in particular on the cases $k = 3$ and $k = 4$ which are amenable to Fourier analysis and "quadratic Fourier analysis" respectively. Then we discuss the recent extension of these theorems to the prime numbers. There is substantial overlap between this survey and [22].

2. Progressions of length three

We now discuss some proofs of Roth's theorem. We first observe that this theorem can be reformulated in one of two equivalent "finitary" settings: firstly as a statement about subsets of long arithmetic progressions, and secondly as a statement about a large cyclic group.

We need some notation. The interval $[a, b]$ shall always refer to the discrete interval $\{n \in \mathbb{Z} : a \leqslant n \leqslant b\}$. We use $|A|$ to denote the cardinality of a finite set A. If A is a finite set and $f : A \to \mathbb{C}$ is a complex-valued function, we define the *expectation* $\mathbb{E}(f) = \mathbb{E}(f(n)|n \in A)$ of f to be the quantity

$$\mathbb{E}(f(n)|n \in A) := \frac{1}{|A|} \sum_{n \in A} f(n);$$

similarly, if $P(n)$ is a property pertaining to elements of A, we define the *probability* of P to be

$$\mathbb{P}(A) = \mathbb{P}(P(n)|n \in A) := \frac{1}{|A|} |\{n \in A : P(n) \text{ is true}\}|,$$

and we define 1_P to be the indicator function of P, thus $1_P(n) = 1$ when $P(n)$ is true and $1_P(n) = 0$ otherwise.

Theorem 2.1 (Roth's theorem, second version)**.** *Let $0 < \delta \leqslant 1$. Then there exists an $N_0 := N_0(\delta) > 1$ such that, for any arithmetic progression $P \subset \mathbb{Z}$ of length at least N_0 and any subset $A \subset P$ of density $\mathbb{P}(n \in A : n \in P) \geqslant \delta$, A contains at least one arithmetic progression $n, n + r, n + 2r$ of length three.*

Note that the choice of progression P is unimportant to this theorem; only the length is relevant. This is because all progressions of a fixed length are isomorphic to each other by an affine scaling map. Thus one could set $P = [1, N]$ here for some $N \geqslant N_0$ with no loss of generality.

Henceforth let us call a function $f : A \to \mathbb{C}$ on a finite set A *bounded* if $|f(n)| \leqslant 1$ for all $n \in A$.

Theorem 2.2 (Roth's theorem, third version)**.** *Let $0 < \delta \leqslant 1$, and let $N \geqslant 1$ be a prime integer. Let $f : \mathbb{Z}/N\mathbb{Z} \to [0, 1]$ be a non-negative bounded function with large mean*

$$\mathbb{E}(f(n)|n \in \mathbb{Z}/N\mathbb{Z}) \geqslant \delta. \tag{2.1}$$

Then we have

$$\mathbb{E}(f(n)f(n + r)f(n + 2r)|n, r \in \mathbb{Z}/N\mathbb{Z}) \geqslant c(3, \delta) - o_\delta(1) \tag{2.2}$$

for some $c(3, \delta) > 0$ depending only on δ, where $o_\delta(1)$ is a quantity that depends on δ and N, and for each fixed δ tends to zero as N goes to infinity.

Before we prove any of these versions, let us first sketch why they are equivalent.

Proof. [Second version implies first version] Let A be a set of positive upper density. Then there exists a $\delta > 0$ such that $|A \cap [1, N]| \geqslant 2\delta N$ for infinitely many N. Using this, one can find infinitely many disjoint intervals $[a_j, b_j]$ of length $b_j - a_j \geqslant N_0(\delta)$ such that A has density at least δ on these intervals:

$$\mathbb{P}(n \in A : n \in [a_j, b_j]) \geqslant \delta.$$

Applying the second version of Roth's theorem to each such interval we thus see A has infinitely many progressions of length 3 as desired. □

Proof. [First version implies second version] Suppose for contradiction that the second version failed. Then we could find a $\delta > 0$ and sets $A_j \subset [1, N_j]$ (with $N_j \to \infty$) with $\mathbb{P}(n \in A_j : n \in [1, N_j]) \geqslant \delta$ and with each A_j containing no arithmetic progressions of length 3. By refining the sequence if necessary we may assume that the N_j are increasing in j (indeed we could make this sequence grow incredibly fast if desired). If one then considers the set $A := \bigcup_{j=1}^\infty 2N_j + A_j$, then it is easy to show that A has positive upper density but contains no arithmetic progressions, a contradiction. $\square$

Proof. [Third version implies second version] Let p be a prime between $2N$ and $4N$ (which always exists by Bertrand's postulate). Let $\pi : [1,N] \to \mathbb{Z}/p\mathbb{Z}$ be the canonical injection of $[1,N]$ into $\mathbb{Z}/p\mathbb{Z}$. If $A \subset [1,N]$ has density $\mathbb{P}(n \in A : n \in [1,N]) \geqslant \delta$, then the function $f := 1_{\pi(A)}$ on $\mathbb{Z}/p\mathbb{Z}$ is non-negative, bounded, and obeys the estimate

$$\mathbb{E}(f(n)|n \in \mathbb{Z}/p\mathbb{Z}) = \frac{|A|}{p} \geqslant \frac{|A|}{4N} \geqslant \delta/4.$$

Thus by the third version of Roth's theorem we have

$$\mathbb{E}(f(n)f(n+r)f(n+2r)|n,r \in \mathbb{Z}/p\mathbb{Z}) \geqslant c(3,\delta/4) - o_\delta(1).$$

Note that $f(n)f(n+r)f(n+2r)$ is non-zero only when $n = \pi(n')$, $n+r = \pi(n'+r')$, $n+2r = \pi(n'+2r')$ and $n' \in [1,N]$, $-N < r < N$, in which case this quantity is equal to 1. Thus we have

$$|\{(n',r') : n', n'+r', n'+2r' \in A; n' \in [1,N]; -N < r' < N\}| \geqslant c(3,\delta/4)p^2 - o_\delta(p^2).$$

We can discard the $r' = 0$ terms as they contribute $O(N) = o(p^2)$. By symmetry we can then reduce to the positive r'. We thus have

$$|\{(n',r') : n', n'+r', n'+2r' \in A; n' \in [1,N]; 0 < r' < N\}| \geqslant c(3,\delta/4)p^2/2 - o_\delta(p^2).$$

If N (and hence p) is sufficiently large, then the right-hand side is non-zero, and we have demonstrated the existence of a non-trivial arithmetic progression of length three in A. (In fact we have demonstrated $\geqslant c'(3,\delta)N^2$ such progressions for some $c'(3,\delta) > 0$). $\square$

Proof. [Second version implies third version] This argument is due to Varnavides [45]. We first observe that to prove the theorem, it suffices to do so when f is a characteristic function $f = 1_A$. This is because if f is non-negative, bounded and obeys (2.1) then the set $A := \{n \in \mathbb{Z}/N\mathbb{Z} : f(n) \geqslant \delta/2\}$ must have density at least $\mathbb{P}(n \in A : n \in \mathbb{Z}/N\mathbb{Z}) \geqslant \delta/2)$. Since we have the pointwise bound[1] from below $f \geqslant \frac{\delta}{2} 1_A$, we have

$$\mathbb{E}(f(n)f(n+r)f(n+2r)|n,r \in \mathbb{Z}/N\mathbb{Z}) \geqslant \frac{\delta^3}{8}\mathbb{E}(1_A(n)1_A(n+r)1_A(n+2r)|n,r \in \mathbb{Z}/N\mathbb{Z})$$

and so (2.2) for f would follow from (2.2) for A (with a slightly worse value of $c(3,\delta)$, namely $\frac{\delta^3}{8}c(3,\delta/2)$).

It remains to verify (2.2) for characteristic functions. Let $M = M(\delta)$ be a large integer depending on δ to be chosen later. To prove (2.2) it suffices to do so in the case $N \gg M$, since the case $N = O(M)$ is vacuous.

The idea is to cover $\mathbb{Z}/N\mathbb{Z}$ uniformly by progressions $P_{ab} := \{a+b, a+2b, \ldots, a+Mb\}$ of length M, where we allow b to be zero. Indeed we observe that for every $n \in \mathbb{Z}/N\mathbb{Z}$ there are exactly NM pairs $(a,b) \in \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$ such that $n \in a + [1,M] \cdot b$ (this is easiest to see by choosing b first). Thus

$$\begin{aligned}\delta &\leqslant \mathbb{P}(n \in A | n \in \mathbb{Z}/N\mathbb{Z}) \\ &= \mathbb{P}(n \in A | n \in P_{ab}; (a,b) \in \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}) \\ &= \mathbb{E}(\mathbb{P}(n \in A | n \in P_{ab} | (a,b) \in \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}).\end{aligned}$$

[1]This is somewhat crude. A slightly better argument would be to select A randomly, with each element $n \in \mathbb{Z}/N\mathbb{Z}$ having a probability of $f(n)$ to lie in A, and then take averages, but in practice this does not yield significantly better constants at the end.

In particular, if we let $\Omega \subseteq \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$ be the set of pairs (a,b) such that $\mathbb{P}(n \in A | n \in P_{ab}) \geqslant \delta/2$, then we have

$$\mathbb{P}((a,b) \in \Omega | (a,b) \in \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}) \geqslant \delta/2. \tag{2.3}$$

Now choose $M := N_0(\delta/2)$. From the definition of Ω and the second form of Roth's theorem, we see that for every $(a,b) \in \Omega$, the set $A \cap P_{ab}$ contains at least one non-trivial arithmetic progression $n, n+r, n+2r$ of length three. In particular we have

$$\mathbb{P}(n, n+r, n+2r \in A | n, n+r, n+2r \in P_{ab}; r \neq 0) \geqslant M^{-2}$$

since the number of progressions $n, n+r, n+2r$ in P_{ab} is at most M^2.

Now observe that every progression $n, n+r, n+2r \in \mathbb{Z}/N\mathbb{Z}$ with $r \neq 0$ is contained in exactly the same number of progressions P_{ab}, since they are all isomorphic using affine scaling maps (here we use that $N = |\mathbb{Z}/N\mathbb{Z}|$ is prime). Thus we have

$$\mathbb{P}(n, n+r, n+2r \in A | n, n+r, n+2r \in \mathbb{Z}/N\mathbb{Z}; r \neq 0) \geqslant M^{-2}$$

In particular (adding in the $r = 0$ case) we have

$$\mathbb{E}(\prod_{j=0}^{2} 1_A(x+jr) | x \in \mathbb{Z}/N\mathbb{Z}; r \in \mathbb{Z}/N\mathbb{Z}) \geqslant M^{-2} - o(1)$$

which gives (2.2) as desired (with $c(3,\delta) = M^{-2} = N_0(\delta/2)^{-2}$ for characteristic functions, and hence $c(3,\delta) = \frac{\delta^3}{8} N_0(\delta/4)^{-2}$ for arbitrary functions). □

In light of these equivalent formulations, it is natural to introduce the Lebesgue spaces $L^p(\mathbb{Z}/N\mathbb{Z})$ for $1 \leqslant p \leqslant \infty$, defined as the complex-valued functions on $\mathbb{Z}/N\mathbb{Z}$ equipped with the norm

$$\|f\|_{L^p(\mathbb{Z}/N\mathbb{Z})} := \mathbb{E}(|f|^p)^{1/p} = (\frac{1}{N} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} |f(n)|^p)^{1/p}$$

and to introduce the trilinear form $\Lambda_3 : L^{p_1}(\mathbb{Z}/N\mathbb{Z}) \times L^{p_2}(\mathbb{Z}/N\mathbb{Z}) \times L^{p_3}(\mathbb{Z}/N\mathbb{Z}) \to \mathbb{C}$ by

$$\Lambda_3(f,g,h) := \mathbb{E}(f(n)g(n+r)h(n+2r) | n, r \in \mathbb{Z}/N\mathbb{Z}). \tag{2.4}$$

Here we always assume N to be a large prime (in particular, it is odd). Thus the third version of Roth's theorem can be reformulated as follows: if $f \in L^\infty(\mathbb{Z}/N\mathbb{Z})$ is a non-negative function obeying the bounds

$$0 < \delta \leqslant \|f\|_{L^1(\mathbb{Z}/N\mathbb{Z})} \leqslant \|f\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})} \leqslant 1$$

then

$$\Lambda_3(f,f,f) \geqslant c(3,\delta) - o_\delta(1) \tag{2.5}$$

for some $c(3,\delta) > 0$. Note that the task here is to obtain *lower bounds* on the form $\Lambda_3(f,f,f)$ rather than upper bounds, which are considerably easier to obtain. For instance, from multilinear interpolation (or Young's inequality) it is easy to establish the upper bounds

$$|\Lambda_3(f,g,h)| \leqslant \|f\|_{L^p} \|g\|_{L^q} \|h\|_{L^r} \tag{2.6}$$

whenever $1 \leqslant p,q,r \leqslant \infty$ and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} \leqslant 2$; here f, g, h are arbitrary complex-valued functions. Note that the non-negativity of f and of Λ_3 (i.e. $\Lambda_3(f,g,h)$ is non-negative whenever f, g, h are non-negative) is crucial, since without this one could not even obtain the trivial bound[2] $\Lambda_3(f, f, f) \geqslant 0$, let alone (2.5).

At first glance it does not appear that upper bounds such as (2.6) are useful for proving lower bounds of the type (2.5). However, one can use the multilinearity of Λ_3 to convert upper bounds to lower bounds as follows. Without loss of generality we may take $\mathbb{E}(f)$ to be equal to δ (since if $\mathbb{E}(f) > \delta$ we may simply decrease f and hence $\Lambda_3(f, f, f)$. We decompose[3] f into a "good function" $g := \mathbb{E}(f) = \delta$ and a "bad function" $b := f - \mathbb{E}(f)$, and then we can split $\Lambda_3(f, f, f)$ into eight components:

$$\Lambda_3(f, f, f) = \Lambda_3(g, g, g) + \ldots + \Lambda_3(b, b, b).$$

The first term can be computed explicitly, and can be viewed as a main term:

$$\Lambda_3(g, g, g) = \Lambda_3(\delta, \delta, \delta) = \delta^3.$$

Thus if one can obtain *upper* bounds on the magnitude of the remaining seven terms which add up to less than δ^3, then one can hope to prove (2.5). The bound (2.6) turns out to be too weak to do this, unless δ is very close to 1 (e.g. if $\delta > 2/3$); however, one can do better by replacing the Lebesgue norms with some additional norms, based on the *Fourier transform*

$$\hat{f}(\xi) := \mathbb{E}(f(x) e_N(-x\xi) | x \in \mathbb{Z}/N\mathbb{Z}),$$

where $e_N : \mathbb{Z}/N\mathbb{Z} \to S^1$ is the character $e_N(x) := \exp(2\pi i x/N)$. From the Fourier inversion formula

$$f(x) = \sum_{\xi \in \mathbb{Z}/N\mathbb{Z}} \hat{f}(\xi) e_N(x\xi)$$

we see that

$$\Lambda_3(f, g, h) = \sum_{\xi_1, \xi_2, \xi_3 \in \mathbb{Z}/p\mathbb{Z}} \hat{f}(\xi_1)\hat{g}(\xi_2)\hat{f}(\xi_3)\mathbb{E}(e_N(n\xi_1 + (n+r)\xi_2 + (n+2r)\xi_3) | n, r \in \mathbb{Z}/p\mathbb{Z}).$$

The expectation on the right-hand side equals 1 when $\xi_1 = \xi_3$ and $\xi_2 = -2\xi_1$, and equal to zero otherwise. Thus we have the identity

$$\Lambda_3(f, g, h) = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} \hat{f}(\xi)\hat{g}(-2\xi)\hat{h}(\xi).$$

From the Plancherel identity

$$\|f\|_{L^2(\mathbb{Z}/N\mathbb{Z})} = \|\hat{f}\|_{l^2(\mathbb{Z}/N\mathbb{Z})}$$

and Hölder's inequality, we thus have the estimate

$$|\Lambda_3(f, g, h)| \leqslant \|f\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \|g\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \|\hat{h}\|_{l^\infty(\mathbb{Z}/N\mathbb{Z})} \tag{2.7}$$

and similarly for permutations. We also have the variant

$$|\Lambda_3(f, g, h)| \leqslant \|f\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \|\hat{g}\|_{l^4(\mathbb{Z}/N\mathbb{Z})} \|\hat{h}\|_{l^4(\mathbb{Z}/N\mathbb{Z})} \tag{2.8}$$

This leads to the following criterion to ensure $\Lambda_3(f, f, f)$ is positive.

[2]There is also the slightly better trivial bound $\Lambda_3(f, f, f) \geqslant \|f\|^3_{L^3(\mathbb{Z}/N\mathbb{Z})}/N$ coming from the $r = 0$ term in (2.4), but this lower bound is $o(1)$ and is thus not significantly better than the trivial bound of 0.

[3]This is of course a very simple decomposition. Later on we shall use more sophisticated decompositions, which can be viewed as "arithmetic" versions of the Calderón-Zygmund decomposition in harmonic analysis.

Proposition 2.3. *Let $f \in L^\infty(\mathbb{Z}/N\mathbb{Z})$ have a decomposition of the form $f = g + b$, where*

$$\|g\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})}, \|b\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})} = O(1); \quad \|g\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})}, \|b\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})} = O(\delta). \tag{2.9}$$

Then we have the estimates

$$\Lambda_3(f, f, f) = \Lambda_3(g, g, g) + O(\delta \|\hat{b}\|_{l^\infty(\mathbb{Z}/N\mathbb{Z})}) \tag{2.10}$$

and

$$\Lambda_3(f, f, f) = \Lambda_3(g, g, g) + O(\delta^{5/4} \|\hat{b}\|_{l^4(\mathbb{Z}/N\mathbb{Z})}).$$

Remark 2.4. *Interestingly, estimates of this type (after being suitably localized in phase space) have proven to be crucial in recent progress in understanding the bilinear Hilbert transform (see e.g.* [30]*), or at least in understanding the contribution of individual "trees" to that transform. Indeed there is some formal similarity between the trilinear form Λ_3 and the trilinear form $\Lambda(f, g, h) := p.v. \int\int f(x + t)g(x - t)h(x)\frac{dxdt}{t}$ associated to the bilinear Hilbert transform.*

Proof. From the hypotheses we have

$$\|g\|_{L^2(\mathbb{Z}/N\mathbb{Z})}, \|b\|_{L^2(\mathbb{Z}/N\mathbb{Z})} = O(\delta^{1/2})$$

and hence by Plancherel

$$\|\hat{g}\|_{l^2(\mathbb{Z}/N\mathbb{Z})}, \|\hat{b}\|_{l^2(\mathbb{Z}/N\mathbb{Z})} = O(\delta^{1/2}).$$

On the other hand, from the L^1 bounds on g and b we have

$$\|\hat{g}\|_{l^\infty(\mathbb{Z}/N\mathbb{Z})}, \|\hat{b}\|_{l^\infty(\mathbb{Z}/N\mathbb{Z})} = O(\delta)$$

and so by Hölder's inequality

$$\|\hat{g}\|_{l^4(\mathbb{Z}/N\mathbb{Z})}, \|\hat{b}\|_{l^4(\mathbb{Z}/N\mathbb{Z})} = O(\delta^{3/4}).$$

The claims now follow by decomposing $\Lambda_3(f, f, f)$ into eight pieces as before, setting aside $\Lambda_3(g, g, g)$ as a main term, and using (2.7), (2.8) (and permutations thereof) to estimate all the remaining pieces (which involve at least one copy of b). □

This suggests the following strategy: in order to obtain a non-trivial lower bound on $\Lambda_3(f, f, f)$, we should obtain a splitting $f = g + b$ obeying the bounds (2.9) where the "good" function g already has a large value of $\Lambda_3(g, g, g)$ (thus we shall presumably want g to be non-negative), and the "bad" function b has a small Fourier transform, either in l^∞ norm or l^4 norm. Note that up to polynomial factors of δ, the two norms are somewhat equivalent, as one can easily establish the estimates

$$\|\hat{b}\|_{l^\infty} \leqslant \|\hat{b}\|_{l^4} \leqslant \|\hat{b}\|_{l^\infty}^{1/2} \|\hat{b}\|_{l^2}^{1/2} \leqslant C\delta^{1/4} \|\hat{b}\|_{l^\infty}^{1/2}. \tag{2.11}$$

In the original arguments involving Roth's theorem, the l^∞ norm on the Fourier coefficients was used, but as we shall see later, it is the l^4 norm which is easier to generalize to "higher order" Fourier analysis, which will be necessary to treat the $k \geqslant 4$ case. Let us rather informally call a function b which obeys bounds such as (2.9) *linearly uniform* if the Fourier transform $\hat{b}$ is very small in either l^∞ or l^4; we see from (2.11) that it is not terribly important which norm we choose here. The reason for this terminology is that a linearly uniform function b is one which is uniformly distributed with respect to linear phase functions $e_N(x\xi)$, in the sense that the inner product of b with such functions is small. (This rather vague statement can be made more precise using Weyl's criterion for uniform distribution).

We have already indicated one such candidate for a decomposition, namely the decomposition $f = g+b$ into the expectation $g := \mathbb{E}(f)$ and the expectation-free $b := f - \mathbb{E}(f)$ components of f. Certainly this decomposition obeys the bounds (2.9), and the value of $\Lambda_3(g, g, g) \geqslant \delta^3$ is moderately large. However, at this stage we do not have very good bounds on $\|\hat{b}\|_{l^\infty(\mathbb{Z}/N\mathbb{Z})}$ or $\|\hat{b}\|_{l^4(\mathbb{Z}/N\mathbb{Z})}$; the best bounds we have on these quantities are $O(\delta)$ and $O(\delta^{3/4})$ respectively, and thus the error term can dominate the main term. (Indeed, there certainly exist functions f for which $\Lambda_3(f, f, f)$ is significantly different from $\Lambda_3(g, g, g)$; consider for instance $f = 1_{[1,\delta N]}$, in which the former quantity is comparable to δ^2 and the latter is comparable to δ^3).

However, we can at least eliminate one case, in which $b = f - \mathbb{E}(f)$ is sufficiently linearly uniform (for instance if $\|\hat{b}\|_\infty \leqslant \delta^2/100$). The question is then what to do in the remaining cases, when $b = f - \mathbb{E}(f)$ is not sufficiently linear uniform. The strategy is then to *convert* the lack of linear uniformity from a liability to an asset, by showing that this lack of uniformity implies some additional structure which one can exploit to improve the situation. The known proofs of Roth's theorem (or more generally Szemerédi's theorem) differ on exactly what this additional structure could be, and how to exploit it, but they essentially fall into one of two categories[4]:

- A *density increment argument* seeks to use the lack of uniformity in b to pass from $\mathbb{Z}/N\mathbb{Z}$ (or $[1, N]$) to a smaller object on which the function f (or the set A) has a larger density. One then iterates this procedure until uniformity is obtained; this algorithm terminates since the density is bounded.
- An *energy increment argument* seeks to use the lack of uniformity in b to improve the decomposition $f = g+b$, replacing the good function g by a function of larger energy (L^2 norm). One then iterates this procedure until uniformity is obtained; this algorithm terminates since the energy is bounded.

Both approaches are important to the theory, as they have different strengths and weaknesses. We illustrate this by giving two proofs of Roth's theorem, one for each of the above approaches. But we shall need some additional notation first; this notation may seem somewhat cumbersome for this application, but will become very convenient when we discuss the case of larger k in later sections.

Definition 2.5 (σ-algebras)**.** Let X be a finite set (such as $\mathbb{Z}/N\mathbb{Z}$ or $[1, N]$). A *σ-algebra* $\mathcal{B}$ in X is any collection of subsets of X which contains the empty set $\emptyset$ and the full set X, and is closed under complementation, unions and intersections. We define the *atoms* of a σ-algebra to be the minimal non-empty elements of $\mathcal{B}$ (with respect to set inclusion); it is clear that the atoms in $\mathcal{B}$ form a partition of X, and $\mathcal{B}$ consists precisely of arbitrary unions of its atoms (including the empty union $\emptyset$); thus there is a one-to-one correspondence between σ-algebras and partitions of X. A function $f : X \to \mathbb{C}$ is said to be *measurable* with respect to a σ-algebra $\mathcal{B}$ if all the level sets of f lie in $\mathcal{B}$, or equivalently if f is constant on each of the atoms of $\mathcal{B}$. We define $L^2(\mathcal{B})$ be the space of $\mathcal{B}$-measurable functions, equipped with the Hilbert space inner product $\langle f, g\rangle_{L^2(X)} := \mathbb{E}(f\overline{g})$. We can then define the conditional expectation operator $f \mapsto \mathbb{E}(f|\mathcal{B})$ to be the orthogonal projection of $L^2(X)$ to $L^2(\mathcal{B})$. An equivalent definition

[4]Szemerédi's proof of Szemerédi's theorem in [39] is a blend of the density increment and energy increment arguments.

of conditional expectation is

$$\mathbb{E}(f|\mathcal{B})(x) := \mathbb{E}(f(y)|y \in \mathcal{B}(x))$$

for all $x \in X$, where $\mathcal{B}(x)$ is the unique atom in $\mathcal{B}$ which contains x. It is clear that conditional expectation is a linear self-adjoint orthogonal projection on $L^2(\mathbb{Z}_N)$, preserves non-negativity, expectation, and constant functions. In particular it maps bounded functions to bounded functions. If $\mathbb{E}(f|\mathcal{B})$ is zero we say that f is *orthogonal to* $\mathcal{B}$.

If $\mathcal{B}$, $\mathcal{B}'$ are two σ-algebras, we use $\mathcal{B} \vee \mathcal{B}'$ to denote the σ-algebra generated by $\mathcal{B}$ and $\mathcal{B}'$ (i.e. the σ-algebra whose atoms are the intersections of atoms in $\mathcal{B}$ with atoms in $\mathcal{B}'$).

Proof. [Density increment proof of Roth's theorem] We now give what is essentially Roth's original argument, though not using Roth's original language (in particular, we give the sigma algebras of Bohr sets significantly more prominence in the argument).

It is more convenient to work with the second formulation of Roth's theorem. Let $\delta > 0$, and let N be a sufficiently large number depending on δ. Let P_0 be a progression of length N, and let A be a subset of P_0 of density at least δ. Our task is to prove that A contains at least one arithmetic progression.

Without loss of generality we may take $P_0 = [1, N]$. Set $\delta_0 := \mathbb{P}(n \in A : n \in [1, N])$, thus $\delta \leqslant \delta_0 \leqslant 1$.

Choose a prime p between $2N$ and $4N$. We embed $[1, N]$ into $\mathbb{Z}/p\mathbb{Z}$ in the obvious manner, thus identifying A with a subset of $\mathbb{Z}/p\mathbb{Z}$, of density at least $\delta/4$. Let us let $f : \mathbb{Z}/p\mathbb{Z} \to \mathbb{R}$ be defined by setting $f(x) := 1_A(x)$ when $x \in [1, N]$ and $f(x) = \delta_0$ otherwise; observe that $\mathbb{E}(f) = \delta_0$ by construction. We then split $f = g + b$, where $g := \mathbb{E}(f) \geqslant \delta_0$ and $b := f - \mathbb{E}(f) = 1_A - \delta_0 1_{[1,N]}$.

There are two cases, depending on whether b is linearly uniform or not. Suppose first that b is linearly uniform in the sense that $\|\hat{b}\|_{l^\infty(\mathbb{Z}/p\mathbb{Z})} \leqslant c\delta^2$ for some small absolute constant $0 < c \ll 1$; this is the "easy case". Since $\Lambda_3(g, g, g) = \mathbb{E}(f)^3 \geqslant \delta_0^3 \geqslant \delta^3$, we see from (2.10) that $\Lambda_3(f, f, f) \geqslant c'\delta^2$ for some absolute constant $c' > 0$ (if c is chosen sufficiently small). By definition of f and Λ_3, this means that

$$\mathbb{P}(n, n + r, n + 2r \in A|(n, r) \in \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}) \geqslant c'\delta^2.$$

The contribution of the $r = 0$ case is at most $O(1/p) = O(1/N)$. Thus if N is large enough, we thus see that there exists at least one pair $(n, r) \in \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ with $r \neq 0$ such that $n, n+r, n+2r$ in A. Since $A \subseteq [1, N]$, this forces $n \in [1, N]$ and $1 \leqslant |r| \leqslant N$. Since $p > 2N$, this implies that A (thought now as a subset of $\mathbb{Z}$ rather than $\mathbb{Z}/p\mathbb{Z}$) also contains a non-trivial arithmetic progression $n, n + r, n + 2r$, as claimed.

Now suppose we are in the "hard case" where b is not linearly uniform, then there exists a frequency $\xi \in \mathbb{Z}/p\mathbb{Z}$ such that $|\hat{b}(\xi)| \geqslant c\delta^2$. By definition of b and the Fourier transform, we thus have

$$|\mathbb{E}((1_A(n) - \delta_0 1_{[1,N]}(n))e_p(-n\xi)|n \in \mathbb{Z}/p\mathbb{Z})| \geqslant c\delta^2.$$

Transferring this back from $\mathbb{Z}/p\mathbb{Z}$ to $[1, N]$, we obtain

$$|\mathbb{E}((1_A(n) - \delta_0)e_p(-n\xi)|n \in [1, N])| \geqslant c\delta^2$$

(with a slightly different constant c). If we let $\chi : [1,N] \to \mathbb{C}$ be the linear phase function $\chi(n) := e_p(n\xi)$, we see that $1_A - \delta_0$ thus has some correlation with χ:

$$|\langle 1_A - \delta_0, \chi\rangle_{L^2([1,N])}| \geqslant c\delta^2. \tag{2.12}$$

Now let $0 < \varepsilon \ll 1$ be a small quantity depending on δ to be chosen later. We partition the complex plane $\mathbb{C} = \bigcup_{Q\in\mathbb{Q}_\varepsilon} Q$ into squares of side-length ε in the standard manner (i.e. the corners of the square lie in the lattice $\varepsilon\mathbb{Z}^2$), and let $\mathcal{B}_{\varepsilon,\chi}$ be the σ-algebra on $[1,N]$ generated by the atoms $\{\chi^{-1}(Q) : Q \in \mathbb{Q}_\varepsilon\}$; sets of this type are also known as *Bohr sets*. Observe that there are only $O(1/\varepsilon)$ non-empty atoms. Then on each atom, χ can only vary by at most $O(\varepsilon)$, and thus we have the pointwise estimate

$$\chi - \mathbb{E}(\chi|\mathcal{B}_{\varepsilon,\chi}) = O(\varepsilon).$$

Since $1_A - \delta_0$ is bounded, we thus see from (2.12) and the triangle inequality that

$$|\langle 1_A - \delta_0, \mathbb{E}(\chi|\mathcal{B}_{\varepsilon,\chi})\rangle_{L^2([1,N])}| \geqslant c\delta^2 - O(\varepsilon).$$

Since conditional expectation is self-adjoint, we have

$$\langle 1_A - \delta_0, \mathbb{E}(\chi|\mathcal{B}_{\varepsilon,\chi})\rangle_{L^2([1,N])} = \langle \mathbb{E}(1_A - \delta_0|\mathcal{B}_{\varepsilon,\chi}), \chi\rangle_{L^2([1,N])},$$

and thus by boundedness of χ

$$\|\mathbb{E}(1_A - \delta_0|\mathcal{B}_{\varepsilon,\chi})\|_{L^1([1,N])} \geqslant c\delta^2 - O(\varepsilon).$$

If we choose $\varepsilon := c'\delta^2$ for some suitably small absolute constant $0 < c' \ll 1$, the left-hand side is at least $c\delta^2/2$. Now observe that $\mathbb{E}(1_A - \delta_0|\mathcal{B}_{\varepsilon,\chi})$ has mean zero:

$$\mathbb{E}(\mathbb{E}(1_A - \delta_0|\mathcal{B}_{\varepsilon,\chi})) = \mathbb{E}(1_A - \delta_0) = \mathbb{E}(1_A) - \delta_0 = \delta_0 - \delta_0 = 0.$$

Thus we see that the positive part of $\mathbb{E}(1_A - \delta_0|\mathcal{B}_{\varepsilon,\chi})$ is large:

$$\mathbb{E}(\mathbb{E}(1_A - \delta_0|\mathcal{B}_{\varepsilon,\chi})_+) \geqslant c\delta^2/4.$$

Now recall that $\mathcal{B}_{\varepsilon,\chi}$ is generated by $O(1/\varepsilon) = O(\delta^{-2})$ non-empty atoms. By definition of conditional expectation and the pigeonhole principle, we can thus find some atom $\chi^{-1}(Q)$ of $\mathcal{B}_{\varepsilon,\chi}$ of density at least $c''\delta^4$ such that $1_A - \delta_0$ is biased on this atom:

$$\mathbb{E}(1_A(n) - \delta_0|n \in \chi^{-1}(Q)) \geqslant c'''\delta^2,$$

and thus

$$\mathbb{P}(n \in A|n \in \chi^{-1}(Q)) \geqslant \delta_0 + c'''\delta^2. \tag{2.13}$$

This is a *density increment*; A is denser on $\chi^{-1}(Q)$ than it is on $[1,N]$. However, $\chi^{-1}(Q)$ is a Bohr set instead of an arithmetic progression. However, the Bohr set is in some sense "very close" to an arithmetic progression in the sense that it can be covered quite efficiently by somewhat long arithmetic progressions[5]. This can be seen as follows. By the pigeonhole principle, one can find an integer $1 \leqslant q \leqslant \sqrt{N}$ such that

$$\|q\frac{\xi}{p}\| \leqslant \frac{1}{\sqrt{N}},$$

where $\|x\|$ denotes the distance of x to the nearest integer. From this one easily observes that if $n \in \chi^{-1}(Q)$, then there is an arithmetic progression containing n of spacing q and length comparable to $\varepsilon\sqrt{N}$ which is completely contained in $\chi^{-1}(Q)$. In particular, one

[5]This step is not particularly efficient when it comes to quantitative constants. A more refined argument of Bourgain [5] works entirely with Bohr sets rather than arithmetic progressions, and obtains the best bounds on $N_0(\delta)$ to date (namely $N_0(\delta) \leqslant C\delta^{-C/\delta^2}$).

can partition $\chi^{-1}(Q)$ into disjoint arithmetic progressions, each of length comparable to $\varepsilon\sqrt{N} \geqslant C^{-1}\delta^2\sqrt{N}$. From (2.13) and the pigeonhole principle, we thus see that at least one of these progressions P_1 has large density:

$$\mathbb{P}(n \in A | n \in P_1) \geqslant \delta_0 + c'''\delta^2.$$

To summarize, we had started with a subset A of a progression P_0 of length N which had density δ_0, and concluded that either A contained an arithmetic progression, or there was a sub-progression P' of length at least $C^{-1}\delta^2\sqrt{N}$ where A has density $\delta_1 > \delta_0 + c'''\delta^2$ for some absolute constant $c''' > 0$. We can then pass to this progression P' and repeat the argument (note that we can make $C^{-1}\delta^2\sqrt{N}$ as large as we please by requiring N to be sufficiently large). The density can only increase by $c'''\delta^2$ by at most $O(1/\delta^2)$ times[6], and so this argument must eventually yield a non-trivial arithmetic of length three in A. □

Proof. [Energy increment proof of Roth's theorem] We now give an energy increment proof of Roth's theorem, inspired by arguments of Furstenberg [10], Bourgain [4], and Green [20], as well as later arguments by Green and the author in [24], [41]. This is not the shortest such proof, nor the most efficient as far as explicit bounds are concerned, but it is a proof which has a relatively small reliance on Fourier analysis and thus which generalizes fairly easily to general k. The structure of this argument, and the concepts introduced, are particularly crucial when establishing long arithmetic progressions in the primes.

We shall use the third formulation of Roth's theorem; unlike the preceding proof, we will not oscillate back and forth between progressions and cyclic groups, but remain in a fixed cyclic group $\mathbb{Z}/N\mathbb{Z}$ throughout. Thus, we let N be a large prime, and let f be a bounded non-negative function on $\mathbb{Z}/N\mathbb{Z}$ obeying the bound (2.1). Our task is to prove (2.5).

We need some additional notation.

Definition 2.6 (Almost periodic functions)**.** A *linear phase function* is a function $\chi : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$ of the form $\chi(n) = e_N(n\xi)$ for some $\xi \in \mathbb{Z}/N\mathbb{Z}$, which we refer to as the *frequency* of χ. If $K > 0$, then an K-*quasiperiodic function* is a function f of the form $\sum_{j=1}^{K} c_j\chi_j$, where each χ_j is a linear phase function (not necessarily distinct), and c_j are scalars such that $|c_j| \leqslant 1$. If $\sigma > 0$, then an (σ, K)-*almost periodic function* is a function f such that $\|f - f_{QP}\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant \sigma$ for some K-quasiperiodic function f_{QP}.

Observe that if f and g are (σ, K)-almost periodic functions, then fg is a $(2\sigma, K^2)$-almost periodic function (taking $(fg)_{K^2} := f_K g_K$).

A key property of almost periodic functions is that one can obtain non-trivial lower bounds on the Λ_3 quantity:

Lemma 2.7 (Almost periodic functions are recurrent)**.** *Let* $0 < \delta < 1$ *and* $0 < \sigma \leqslant \delta^3/100$*, and* f *be an bounded non-negative* (σ, K)*-almost periodic function obeying* (2.1)*. Then we have*

$$\Lambda_3(f, f, f) \geqslant c(K, \delta) - o_{n,\delta}(1)$$

for some $c(K, M, \delta) > 0$ *(the key point here being that this quantity is independent of* N*).*

[6]One can improve this to $O(1/\delta)$ by observing that the density increment of $c'''\delta^2$ can be refined to $c'''\delta_0^2$.

Proof. Let $f_{QP} = \sum_{j=1}^K c_j \chi_j$ be the K-quasiperiodic function approximating f, and let $0 < \varepsilon$ be a small number (depending on K, δ) to be chosen later. Let $\xi_1, \ldots, \xi_K$ be the frequencies associated to the characters $\chi_1, \ldots, \chi_K$. By Dirichlet's simultaneous approximation by rationals theorem (or the pigeonhole principle), we have

$$\mathbb{P}(\|r\xi_j\| \leqslant \varepsilon \text{ for all } 1 \leqslant j \leqslant K | r \in \mathbb{Z}/N\mathbb{Z}) \geqslant c(\varepsilon, K) \tag{2.14}$$

for some $c(\varepsilon, K) > 0$ independent of N. Next, observe from the triangle inequality that if r is as above, then

$$\|T^r f_{AP} - f_{AP}\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant C(K)\varepsilon$$

where T^r is the shift map $T^r f(x) := f(x+r)$. From this and the triangle inequality, we conclude

$$\|T^r f - f\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant \delta^3/10 + C(K)\varepsilon,$$

and by another application of T^r, we have

$$\|T^{2r} f - T^r f\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant \delta^3/10 + C(K)\varepsilon.$$

From this and the boundedness of f, we conclude that

$$\|f T^r f T^{2r} f - f^3\|_{L^1(\mathbb{Z}/N\mathbb{Z})} \leqslant \delta^3/2 + C(K)\varepsilon,$$

but from the bounded non-negativity of f, (2.1), and Hölder's inequality we have

$$\|f^3\|_{L^1(\mathbb{Z}/N\mathbb{Z})} \geqslant \|f\|^3_{L^1(\mathbb{Z}/N\mathbb{Z})} \geqslant \delta^3$$

and hence (by positivity of f)

$$\mathbb{E}(f T^r f T^{2r} f(n) | n \in \mathbb{Z}/N\mathbb{Z}) \geqslant \delta^3/2 - C(K)\varepsilon.$$

If we choose ε small enough depending on δ and M, we thus have

$$\mathbb{E}(f T^r f T^{2r} f(n) | n \in \mathbb{Z}/N\mathbb{Z}) \geqslant \delta^3/4.$$

Averaging over all r, using (2.14) and the non-negativity of f, we obtain

$$\mathbb{E}(f T^r f T^{2r} f(n) | n, r \in \mathbb{Z}/N\mathbb{Z}) \geqslant \delta^3 c(\varepsilon, K)/4.$$

But the left-hand side is nothing more than $\Lambda_3(f, f, f)$. The claim follows. □

To exploit the above result we shall need to approximate a general function f by an almost periodic function, plus a linearly uniform error. The first step in this strategy shall be to construct σ-algebras such that the measurable functions in this algebra are all almost periodic.

Lemma 2.8. *Let $0 < \varepsilon \ll 1$ and let χ be a linear phase function. Then there exists a σ-algebra $\mathcal{B}_{\varepsilon,\chi}$ such that $\|\chi - \mathbb{E}(\chi|\mathcal{B}_{\varepsilon,\chi})\|_{L^\infty} \leqslant C\varepsilon$, and such that for every $\sigma > 0$, there exists $K = K(\sigma, \varepsilon)$ such that every function f which is measurable with respect to $\mathcal{B}_{\varepsilon,\chi}$ and obeys the bound $\|f\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})} \leqslant 1$ is (σ, K)-almost periodic.*

Proof. We use a random construction, constructing a σ-algebra which has the stated properties with non-zero probability. Let α be a randomly selected element of the unit square in the complex plane, and let $\mathcal{B}_{\varepsilon,\chi}$ be the σ-algebra with atoms of the form $\{\chi^{-1}(Q) : Q \in \mathbb{Q}_\varepsilon + \varepsilon\alpha\}$. Then as in the previous proof of Roth's theorem, we have $\|\chi - \mathbb{E}(\chi|\mathcal{B}_{\varepsilon,\chi})\|_{L^\infty} \leqslant C\varepsilon$. Now we prove the approximation claim. It suffices to verify the claim for $\sigma = 2^{-n}$ for some integer $n \gg 1$, with probability $1 - O(\sigma)$. Also, since $\mathcal{B}_{\varepsilon,\chi}$ has at most $C(\varepsilon)$ atoms, it suffices to verify the claim when f is the indicator function of one of these atoms A, with probability $1 - O(C(\varepsilon)^{-1}\sigma)$.

The function f can be rewritten as $f(x) = 1_Q(\chi(x) - \varepsilon\alpha)$. We can use the Weierstrass approximation theorem to approximate $1_Q(z)$ on the disk $z = O(1/\varepsilon)$ by a polynomial $P(z, \overline{z})$ involving at most $C(\sigma, \varepsilon)$ terms and with coefficients bounded by $C(\sigma, \varepsilon)$ such that $|P|$ is bounded by 1 in this disk, and $1_Q(z) - P(z, \overline{z}) = O(C^{-1}\sigma)$ for all z in this disk, except for a set of measure $O(C(\varepsilon)^{-2}\sigma^2)$. A standard randomization argument then allows us to assert that

$$\|1_Q(\chi(x) - \varepsilon\alpha) - P(\chi(x) - \varepsilon\alpha, \overline{\chi(x) - \varepsilon\alpha})\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant \sigma$$

with probability $O(C(\varepsilon)^{-1}\sigma)$. But $P(\chi(x) - \varepsilon\alpha, \overline{\chi(x) - \varepsilon\alpha})$ can be written as the linear combination of at most $C(\varepsilon, \sigma)$ characters, with coefficients at most $C(\varepsilon, \sigma)$, and is thus $C(\varepsilon, \sigma)$-quasiperiodic (one can reduce the coefficients to be less than 1 by repeating characters as necessary). The claim follows. □

One can concatenate these σ-algebras together. If $\mathcal{B}_1, \ldots, \mathcal{B}_n$ are σ-algebras, we let $\mathcal{B}_1 \vee \ldots \vee \mathcal{B}_n$ be the smallest σ-algebra which contains all of them.

Corollary 2.9. *Let $0 < \varepsilon_1, \ldots, \varepsilon_n \ll 1$ and let $\chi_1, \ldots, \chi_n$ be linear phases. Let $\mathcal{B}_{\varepsilon_1,\chi_1}, \ldots, \mathcal{B}_{\varepsilon_1,\chi_n}$ be the σ-algebras arising from the above corollary. Then for every $\sigma > 0$, there exists $K = K(n, \sigma, \varepsilon_1, \ldots, \varepsilon_n)$ such that every function f which is measurable with respect to $\mathcal{B}_{\varepsilon_1,\chi_1} \vee \ldots \vee \mathcal{B}_{\varepsilon_1,\chi_n}$ and obeys the bound $\|f\|_{L^\infty(\mathbb{Z}/N\mathbb{Z})} \leqslant 1$ is (σ, K)-almost periodic.*

Proof. Since the number of atoms in this σ-algebra is at most $C(n, \varepsilon_1, \ldots, \varepsilon_n)$, it suffices to verify this when f is the indicator function of a single atom. But then f is the product of n indicator functions from atoms in $\mathcal{B}_{\varepsilon_1,\chi_1}, \ldots, \mathcal{B}_{\varepsilon_n,\chi_n}$, and the claim follows from the preceding lemma and the previously made observation that the product of almost periodic functions is almost periodic. □

The significance of these σ-algebras is not only that they contain functions which are almost periodic and hence have non-trivial bounds on the Λ_3 form, but also that they capture "obstructions to linear uniformity":

Lemma 2.10 (Non-uniformity implies structure)**.** *Let b be a bounded function such that $\|\hat{b}\|_{l^\infty} \geqslant \sigma > 0$, and let $0 < \varepsilon \ll \sigma$. Then there exists a linear phase function χ with associated σ-algebra $\mathcal{B}_{\varepsilon,\chi}$ such that*

$$\|\mathbb{E}(b|\mathcal{B}_{\varepsilon,\chi})\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \geqslant C^{-1}\sigma.$$

This is proven by a repetition of the arguments used in the first proof of Roth's theorem, and we leave it to the reader.

We can now assemble all these ingredients together to prove Roth's theorem. The major step here is a *structure theorem* which decomposes an arbitrary function into an almost periodic piece and a linearly uniform piece.

Proposition 2.11 (Quantitative Koopman-von Neumann theorem)**.** *Let $F : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be an arbitrary function, let $0 < \delta \leqslant 1$, and let f be any bounded non-negative function on $\mathbb{Z}/N\mathbb{Z}$ obeying* (2.1)*. Let $\sigma := \delta^3/100$. Then there exists a quantity $0 < K \leqslant C(F, \delta)$ and a decomposition $f = g + b$, where g is bounded, non-negative, has mean $\mathbb{E}(g) = \mathbb{E}(f)$, and (σ, K)-almost periodic, and b obeys the bound*

$$\|\hat{b}\|_{l^\infty} \leqslant F(\delta, K). \tag{2.15}$$

Proof. We apply the following *energy incrementation algorithm* to construct g and b. We shall need two auxiliary σ-algebras $\mathcal{B}$ and $\mathcal{B}'$, with $\mathcal{B}'$ always being larger than or equal to $\mathcal{B}$. Also, $\mathcal{B}$ will always be of the form $\mathcal{B} = \mathcal{B}_{\varepsilon_1,\chi_1} \vee \ldots \vee \mathcal{B}_{\varepsilon_n,\chi_n}$ for some n, some $\varepsilon_1, \ldots, \varepsilon_n > 0$, and some $\chi_1, \ldots, \chi_n$, and similarly for $\mathcal{B}'$ (but with different values of n); also we will have the bound

$$\|\mathbb{E}(f|\mathcal{B}')\|^2_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant \|\mathbb{E}(f|\mathcal{B})\|^2_{L^2(\mathbb{Z}/N\mathbb{Z})} + \sigma^2/4 \tag{2.16}$$

or equivalently (by Pythagoras' theorem)

$$\|\mathbb{E}(f|\mathcal{B}') - \mathbb{E}(f|\mathcal{B})\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \leqslant \sigma/2. \tag{2.17}$$

- Step 0: Initialize $\mathcal{B} = \mathcal{B}' = \{0, \mathbb{Z}/N\mathbb{Z}\}$ to be the trivial σ-algebra. Note that (2.16) is trivially true at present.
- Step 1: By construction, we have $\mathcal{B} = \mathcal{B}_{\varepsilon_1,\chi_1} \vee \ldots \vee \mathcal{B}_{\varepsilon_n,\chi_n}$ for some $\varepsilon_1, \ldots, \varepsilon_n > 0$ and linear phase functions $\chi_1, \ldots, \chi_n$. The function $\mathbb{E}(f|\mathcal{B})$ is bounded and measurable with respect to $\mathcal{B}$. By Corollary 2.9 we can thus find K depending on δ, n, $\varepsilon_1, \ldots, \varepsilon_n$ such that $\mathbb{E}(f|\mathcal{B})$ is $(\sigma/2, K)$-almost periodic.
- Step 2: Set $g := \mathbb{E}(f|\mathcal{B}')$ and $b = f - \mathbb{E}(f|\mathcal{B}')$. If $\|\hat{b}\|_{l^\infty} \leqslant F(\delta, K)$ then we terminate the algorithm; otherwise we move on to Step 3.
- Step 3: Since we have not terminated the algorithm, we have $\|\hat{b}\|_{l^\infty} > F(\delta, K)$. Using Lemma 2.10, we can then find $\varepsilon = F(\delta, K)/C$ and a character χ, with associated σ-algebra $\mathcal{B}_{\varepsilon,\chi}$, such that
$$\|\mathbb{E}(b|\mathcal{B}_{\varepsilon,\chi})\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \geqslant C^{-1}F(\delta, K).$$
From the identity
$$\mathbb{E}(b|\mathcal{B}_{\varepsilon,\chi}) = \mathbb{E}(\mathbb{E}(f|\mathcal{B}' \vee \mathcal{B}_{\varepsilon,\chi}) - \mathbb{E}(f|\mathcal{B}')|\mathcal{B}_{\varepsilon,\chi})$$
and Pythagoras's theorem, we thus have
$$\|\mathbb{E}(f|\mathcal{B}' \vee \mathcal{B}_{\varepsilon,\chi}) - \mathbb{E}(f|\mathcal{B}')\|_{L^2(\mathbb{Z}/N\mathbb{Z})} \geqslant C^{-1}F(\delta, K),$$
which by Pythagoras again implies the energy increment
$$\|\mathbb{E}(f|\mathcal{B}' \vee \mathcal{B}_{\varepsilon,\chi})\|^2_{L^2(\mathbb{Z}/N\mathbb{Z})} \geqslant \|\mathbb{E}(f|\mathcal{B}')\|^2_{L^2(\mathbb{Z}/N\mathbb{Z})} + C^{-2}F(\delta, K)^2.$$
- Step 4: We now replace $\mathcal{B}'$ with $\mathcal{B}' \vee \mathcal{B}_{\varepsilon,\chi}$. If we continue to have the property (2.16), thne we return to Step 2. Otherwise, we replace $\mathcal{B}$ with $\mathcal{B}'$ and return to Step 1.

Let us first see why this algorithm terminates. If $\mathcal{B}$ (and hence K) is fixed, then each time we pass through Step 4, the energy $\|\mathbb{E}(f|\mathcal{B}')\|^2_{L^2(\mathbb{Z}/N\mathbb{Z})}$ increases by at least $C^{-2}F(\delta, K)^2$. Thus either we terminate the algorithm, or (2.16) must be violated, within $C\sigma^2/F(\delta, K)^2 = C(F, \delta, K)$ steps. If the latter occurs, then B is replaced by a new σ algebra involving $C(F, \delta, K)$ new characters, with corresponding ε parameters which are bounded from below by $C(F, \delta, K)^{-1}$. This implies that the K quantity associated to $\mathcal{B}$ will be replaced by a quantity of the form $C(F, \delta, K)$. Also, the energy $\|\mathbb{E}(f|\mathcal{B})\|^2_{L^2(\mathbb{Z}/N\mathbb{Z})}$ will have increased by at least $\sigma^2/4$, thanks to the violation of (2.16). On the other hand, since f was assumed bounded, this energy cannot exceed 1. Thus we can change $\mathcal{B}$ at most $O(\sigma^{-2})$ times. Putting all this together we see that ths entire algorithm must terminate in $C(F, \delta)$ steps, and the quantity K will also not exceed $C(F, \delta)$. (Note that these constants can be extremely large, as they will involve iterating F repeatedly; however, the key point is that they do not depend on N).

The claims of the proposition now follow from construction. Note that $\mathbb{E}(f|\mathcal{B})$ is $(\sigma/2, K)$-almost periodic by construction, and hence $g = \mathbb{E}(f|\mathcal{B}')$ will be (σ, K)-almost periodic thanks to (2.17). □

We can now finally prove Roth's theorem. We let $F : \mathbb{R}^+ \to \mathbb{R}^+ \to \mathbb{R}^+$ be a function to be chosen later, and apply the above Proposition to decompose $f = g + b$. By Lemma 2.7 we have

$$\Lambda_3(g, g, g) \geqslant c(K, \delta) - o_{n,\delta}(1)$$

and then by (2.10) and (2.15) we have

$$\Lambda_3(f, f, f) \geqslant c(K, \delta) + O(\delta F(K, \delta)) - o_{n,\delta}(1).$$

By choosing F sufficiently small, we can absorb the second term in the first, thus

$$\Lambda_3(f, f, f) \geqslant c(K, \delta)/2 - o_{n,\delta}(1).$$

Since $K \leqslant C(F, \delta) = C(\delta)$, the claim (2.5) now follows. □

We remark that there are several other proofs of Roth's theorem in the literature, notably Szemerédi's proof based on density increment arguments and extremely large cubes (see [19]), and an argument based on the Szemerédi regularity lemma (which in turn requires energy increment arguments in the proof) in [37]. While these arguments are also important to the theory and both have generalizations to higher k, we will not discuss them here due to lack of space.

3. Interlude on multilinear operators

We will shortly turn our attention to Szemerédi's theorem. Based on the preceding section, it is unsurprising that much of the analysis will revolve around the multilinear form

$$\Lambda_k(f_0, \ldots, f_{k-1}) := \mathbb{E}\Big(\prod_{j=0}^{k-1} f_j(x + jr) | x, r \in \mathbb{Z}/N\mathbb{Z}\Big)$$

for a large prime N. It turns out that to analyze this multilinear form, it is convenient to generalize substantially and consider multilinear expressions of the form

$$\mathbb{E}\Big(K(x) \prod_{j=1}^{d} F_j(x) | x \in \prod_{j=1}^{d} A_j\Big) \tag{3.1}$$

where $d \geqslant 1$ is fixed, $A_1, \ldots, A_d$ are finite non-empty sets, $K : \prod_{j=1}^d A_j \to \mathbb{C}$ is a fixed kernel, $x = (x_1, \ldots, x_d)$, and each $F_i : \prod_{j=1}^d A_j \to \mathbb{C}$ is a bounded function which is independent of the x_i co-ordinate (and thus only depends on the other $n - 1$ co-ordinates).

Henceforth we fix d and $A_1, \ldots, A_d$. Let $\{0, 1\}^d$ be the discrete unit cube. We need the following notation: if $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$ and $x^{(1)} = (x_1^{(1)}, \ldots, x_d^{(1)})$ are elements of $\prod_{j=1}^d A_j$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d) \in \{0, 1\}^d$, then we write $x^{(\varepsilon)} := (x_1^{(\varepsilon_1)}, \ldots, x_d^{(\varepsilon_d)}) \in \prod_{j=1}^d A_j$, and refer to the 2^d-tuple $(x^{(\varepsilon)})_{\varepsilon \in \{0,1\}^d}$ of elements in $\prod_{j=1}^d A_j$ as the *cube* generated by $x^{(0)}$ and $x^{(1)}$; this is a cube in the combinatorial sense rather than the geometric sense. Thus for instance, when $d = 2$, the cube generated by (x, y) and (x', y') is the 4-tuple consisting of (x, y), (x, y'), (x', y), and (x', y').

Now suppose we have a 2^d-tuple of kernels $K^{(\varepsilon)} : \prod_{j=1}^d A_j \to \mathbb{C}$ for each $\varepsilon \in \{0,1\}^d$. We define the *Gowers inner product* $\langle (K^{(\varepsilon)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d}$ to be

$$\langle (K^{(\varepsilon)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d} := \mathbb{E}(\prod_{\varepsilon \in \{0,1\}^d} \mathcal{C}^{|\varepsilon|} K^{(\varepsilon)}(x^{(\varepsilon)}) | x^{(0)}, x^{(1)} \in \prod_{j=1}^d A_j)$$

where $\mathcal{C} f := \overline{f}$ is the conjugation operator, and $|\varepsilon| := \sum_{j=1}^d \varepsilon_j$. By separating the d^{th} co-ordinates of $x^{(0)}$ and $x^{(1)}$, we observe the identity

$$\begin{aligned} \langle (K^{(\varepsilon)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d} = \mathbb{E}(\mathbb{E}(&\prod_{\underline{\varepsilon} \in \{0,1\}^{d-1}} \mathcal{C}^{|\underline{\varepsilon}|} K^{(\underline{\varepsilon},0)}(\underline{x}^{(\underline{\varepsilon})}, y) | y \in A_d) \\ &\mathcal{C}\mathbb{E}(\prod_{\underline{\varepsilon} \in \{0,1\}^{d-1}} \mathcal{C}^{|\underline{\varepsilon}|} K^{(\underline{\varepsilon},1)}(\underline{x}^{(\underline{\varepsilon})}, y) | y \in A_d) | \underline{x}^{(0)}, \underline{x}^{(1)} \in \prod_{j=1}^{d-1} A_j) \end{aligned} \tag{3.2}$$

Applying Cauchy-Schwarz in the variables $\underline{x}^{(0)}, \underline{x}^{(1)}$, we conclude that

$$|\langle (K^{(\varepsilon)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d}| \leqslant \langle (K^{(\underline{\varepsilon},0)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d}^{1/2} \langle (K^{(\underline{\varepsilon},1)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d}^{1/2},$$

where $\underline{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_{d-1})$ are the first $d-1$ co-ordinates of ε; note that (3.2) ensures that the inner products appearing in the right-hand side of the above equation are non-negative reals. Of course one has a similar inequality if we work with the j^{th} co-ordinate instead of the d^{th} co-ordinate for any $1 \leqslant j \leqslant d$. Applying the above Cauchy-Schwarz inequality once in each co-ordinate, we obtain the *Gowers-Cauchy-Schwarz inequality*

$$\langle (K^{(\varepsilon)})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d} \leqslant \prod_{\varepsilon \in \{0,1\}^d} \|K^{(\varepsilon)}\|_{\Box^d} \tag{3.3}$$

where $\|K\|_{\Box^d}$ is the *Gowers cube norm*

$$\|K\|_{\Box^d} := \langle (K)_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d}^{1/2^d}.$$

Again, the identity (3.2) ensures that this norm is non-negative. Using the multilinearity of the Gowers inner product, we then observe for an arbitrary pair K_0, K_1 of kernels that

$$\begin{aligned} \|K_0 + K_1\|_{\Box^d}^{2^d} &= \langle (K_0 + K_1)_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d} \\ &= \sum_{A \subset \{0,1\}^d} \langle (K_{1_{A(\varepsilon)}})_{\varepsilon \in \{0,1\}^d} \rangle_{\Box^d} \\ &\leqslant \sum_{A \subset \{0,1\}^d} \prod_{\varepsilon \in \{0,1\}^d} \|K_{1_{A(\varepsilon)}}\|_{\Box^d} \\ &= (\|K_0\|_{\Box^d} + \|K_1\|_{\Box^d})^{2^d} \end{aligned}$$

which thus yields the *Gowers triangle inequality*

$$\|K_0 + K_1\|_{\Box^d} \leqslant \|K_0\|_{\Box^d} + \|K_1\|_{\Box^d}.$$

Since the Gowers cube norm is clearly homogeneous, we thus see that $\|\cdot\|_{\Box^d}$ is a semi-norm. We will later show that it is in fact a norm when $d \geqslant 2$; when $d = 1$ we have $\|K\|_{\Box^1} = |\mathbb{E}(K(x)|x \in A_1)|$ which is degenerate and thus not a genuine norm.

The significance of the Gowers cube norm to expressions of the form (3.1) lies in the following estimate (which is implicit in [8] and also in [17]).

Lemma 3.1 (Van der Corput lemma)**.** *Let $d \geqslant 1$, let $A_1, \ldots, A_d$ be finite non-empty sets, let $K : \prod_{j=1}^d A_j \to \mathbb{C}$, and for each $1 \leqslant i \leqslant d$ let $F_i : \prod_{j=1}^d A_j \to \mathbb{C}$ is a bounded function which is independent of the x_i co-ordinate. Then we have*

$$|\mathbb{E}(K(x)\prod_{j=1}^d F_j(x)|x \in \prod_{j=1}^d A_j)| \leqslant \|K\|_{\square^d}.$$

Proof. We induct on d. When $d = 1$, the claim becomes

$$|\mathbb{E}(K(x_1)F_1(x_1)|x_1 \in A_1)| \leqslant |\mathbb{E}(K(x_1)|x_1 \in A_1)|,$$

which follows since F_1 is independent of x_1 and is bounded.

Now suppose that $d \geqslant 2$ and the claim has already been proven for $d - 1$. Since F_d is independent of the x_d co-ordinate, we may abuse notation and interpret F_d as a function on $\prod_{j=1}^{d-1} A_j$ rather than $\prod_{j=1}^d A_d$. We then separate off the x_d co-ordinate to write

$$\mathbb{E}(K(x)\prod_{j=1}^d F_j(x)|x \in \prod_{j=1}^d A_j) = \mathbb{E}(F_d(\underline{x})\mathbb{E}(K(\underline{x}, x_d)\prod_{j=1}^{d-1} F_j(\underline{x}, x_d)|x_d \in A_d)|\underline{x} \in \prod_{j=1}^{d-1} A_j).$$

Since F_d is bounded, we may apply Cauchy-Schwarz in the $\underline{x}$ variable to then obtain

$$\begin{aligned}|\mathbb{E}(K(x)\prod_{j=1}^d F_j(x)|x \in \prod_{j=1}^d A_j)| &\leqslant \mathbb{E}(|\mathbb{E}(K(\underline{x}, x_d)\prod_{j=1}^{d-1} F_j(\underline{x}, x_d)|x_d \in A_d)|^2|\underline{x} \in \prod_{j=1}^{d-1} A_j) \\ &= \mathbb{E}(\mathbb{E}(K(\underline{x}, x_d^{(0)})\overline{K(\underline{x}, x_d^{(1)})}\prod_{j=1}^{d-1} F_j(\underline{x}, x_d^{(0)})\overline{F_j(\underline{x}, x_d^{(1)})}|\underline{x} \in \prod_{j=1}^{d-1} A_j) \\ &\quad |x_d^{(0)}, x_d^{(1)} \in A_d)^{1/2}.\end{aligned}$$

For each fixed $x_d^{(0)}, x_d^{(1)} \in A$ and each $1 \leqslant j \leqslant d$, the function $F_j(\underline{x}, x_d^{(0)})\overline{F_j(\underline{x}, x_d^{(1)})}$ is a bounded function of $\underline{x}$. If we then apply the induction hypothesis we have

$$|\mathbb{E}(K(x)\prod_{j=1}^d F_j(x)|x \in \prod_{j=1}^d A_j)| \leqslant \mathbb{E}(\|K(\cdot, x_d^{(0)})\overline{K(\cdot, x_d^{(1)})}\|_{\square^{d-1}}|x_d^{(0)}, x_d^{(1)} \in A_d)^{1/2},$$

so by Hölder's inequality

$$|\mathbb{E}(K(x)\prod_{j=1}^d F_j(x)|x \in \prod_{j=1}^d A_j)| \leqslant \mathbb{E}(\|K(\cdot, x_d^{(0)})\overline{K(\cdot, x_d^{(1)})}\|_{\square^{d-1}}^{2^{d-1}}|x_d^{(0)}, x_d^{(1)} \in A_d)^{1/2^d}.$$

But the right-hand side can be re-arranged to be precisely $\|K\|_{\square^d}$, and the claim follows. □

We can now show that $\|\cdot\|_{U^d}$ is a genuine norm when $d \geqslant 2$:

Corollary 3.2. *If $d \geqslant 2$ and $\|K\|_{U^d} = 0$, then $K = 0$.*

Proof. Let $(x_1, \ldots, x_d) \in \prod_{j=1}^d A_j$ be arbitrary. We then define $f_i : \prod_{j=1}^d A_j \to \mathbb{C}$ by defining $f_i(y_1, \ldots, y_d) = 1$ when $y_j = x_j$ for all $j \neq i$, and $f_i(y_1, \ldots, y_d) = 0$ otherwise. Applying the previous lemma we thus see that $K(x_1, \ldots, x_d) = 0$. Since $(x_1, \ldots, x_d)$ was arbitrary, the claim follows. □

Let us informally call a kernel K *Gowers uniform* if it has small $\Box^d$ norm. Then the van der Corput lemma then asserts that Gowers uniform kernels are negligible for the purpose of computing multilinear expressions such as (3.1). In particular, when $d = 2$, the $\Box^2$ norm of a kernel K (which can now be interpreted as a linear operator T_K from $L^2(A_1)$ to $L^2(A_2)$) controls the L^2 operator norm of K. Indeed, one has the identity

$$\begin{aligned} \|K\|_{U^2} &= \|T_K^* T_K\|_{HS(L^2(A_1)\to L^2(A_1))}^{1/2} \\ &= \|T_K T_K^*\|_{HS(L^2(A_2)\to L^2(A_2))}^{1/2} \\ &= \mathrm{tr}_{A_1}(T_K^* T_K T_K^* T_K)^{1/4} \\ &= \mathrm{tr}_{A_2}(T_K T_K^* T_K T_K^*)^{1/4} \end{aligned} \tag{3.4}$$

where HS is the normalized Hilbert-Schmidt norm, and tr_A is the normalized trace on A; equivalently, $\|K\|_{\Box^2}$ is the l^4 norm of the (normalized) singular values of K, while the operator norm is the l^∞ norm of these singular values (and the Hilbert-Schmidt norm is the l^2 norm). Thus one can view the $\Box^d$ norm as a multilinear generalization of the l^4 Schatten-von Neumann norm. This norm has also arisen in the study of pseudorandom sets and graphs, see for instance [6].

Now we specialize to the problem of counting arithmetic progressions in $\mathbb{Z}/N\mathbb{Z}$.

Definition 3.3 (Gowers uniformity norm)**.** Let $f : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$ be a function and $d \geqslant 1$. Then we define the *Gowers uniformity norm* $\|f\|_{U^d}$ to be the quantity $\|f\|_{U^d} := \|K\|_{\Box^d}$, where $K : (\mathbb{Z}/N\mathbb{Z})^d \to \mathbb{C}$ is the kernel

$$K(x_1, \ldots, x_d) := f(x_1 + \ldots + x_d).$$

Equivalently, we have

$$\|f\|_{U^d} := \mathbb{E}(\prod_{\varepsilon \in \{0,1\}^d} \mathcal{C}^{|\varepsilon|} f(x + \sum_{j=1}^d \varepsilon_j h_j) | x, h_1, \ldots, h_d \in \mathbb{Z}/N\mathbb{Z})^{1/2^d},$$

or alternatively we have the recursive definitions

$$\|f\|_{U^1} := |\mathbb{E}(f)|; \quad \|f\|_{U^{d+1}} := \mathbb{E}(\|f(x+h)\overline{f(x)}\|_{U^d_x}^{2^d} | h \in \mathbb{Z}/N\mathbb{Z})^{1/2^{d+1}}. \tag{3.5}$$

Since $\Box^d$ was a norm for $d \geqslant 2$, we see that U^d is also a norm when $d \geqslant 2$. In the $d = 2$ case, one can easily verify the identity

$$\|f\|_{U^2} := \|\hat{f}\|_{l^4},$$

which can be viewed as a special case of (3.4), observing that the Fourier coefficients of f are essentially the eigenvalues of K. However, for $d \geqslant 3$ the U^d norm becomes more complicated, and has no particularly useful representation in terms of the Fourier transform. Using the Gowers-Cauchy-Schwarz inequality, it is possible to show the monotonicity relationship $\|f\|_{U^d} \leqslant \|f\|_{U^{d+1}}$ for all d; one can also show that $\|f\|_{U^d} \to \|f\|_{L^\infty}$ as $d \to \infty$. We shall neither prove nor use these facts here.

We can now obtain an analogue of (2.8).

Lemma 3.4 (Generalized von Neumann theorem)**.** [17] *Let $k \geqslant 3$, and let N be a prime larger than k. Let $f_0, \ldots, f_{k-1}$ be bounded functions on $\mathbb{Z}/N\mathbb{Z}$. Then we have*

$$|\Lambda_k(f_0, \ldots, f_{k-1})| \leqslant \min_{0 \leqslant j \leqslant k-1} \|f_j\|_{U^{k-1}}.$$

Proof. Fix $0 \leqslant j \leqslant k-1$; it thus suffices to show that

$$|\Lambda_k(f_0, \ldots, f_{k-1})| \leqslant \|f_j\|_{U^{k-1}}.$$

Observe that for any $x_1, \ldots, x_{k-1} \in \mathbb{Z}/N\mathbb{Z}$, the sequence

$$(x_1 + \ldots + x_{k-1} - (j-i) \sum_{1 \leqslant i' \leqslant k : i' \neq j} \frac{1}{j-i'} x_{i'})_{1 \leqslant i \leqslant k}$$

is an arithmetic progression of length k in $\mathbb{Z}/N\mathbb{Z}$ (here we are using the hypothesis that N is prime and larger than k in order to invert $j - i'$). Conversely, each progression $x, x+r, \ldots, x+(k-1)r$ can be expressed in the above form in exactly the same number of ways (N^{k-3}, to be exact). We may thus write

$$\Lambda_k(f_0, \ldots, f_{k-1}) = \mathbb{E}(\prod_{i=0}^{k-1} f_i(x_1 + \ldots + x_{k-1} - (j-i) \sum_{1 \leqslant i' \leqslant k : i' \neq j} \frac{1}{j-i'} x_{i'}) | x_1, \ldots, x_{k-1} \in \mathbb{Z}/N\mathbb{Z}).$$

Now observe that the i^{th} factor in the above sum is bounded and will not depend on x_i when $i \neq j$, and that the j^{th} factor is $f_j(x_1 + \ldots + x_{k-1})$. Applying the van der Corput lemma and the definition of the U^{k-1} norm, we obtain the claim. □

Let us informally call a bounded function f *Gowers uniform of order* $k-2$ if $\|f\|_{U^{k-1}}$ is small; thus for instance a function with small U^2 norm is linearly uniform, a function with small U^3 norm is quadratically uniform, and so forth. The above lemma then asserts that functions which are Gowers uniform of order $k-2$ have a negligible impact on the Λ_k multilinear form.

Example 3.5. *Let N be a prime number, let $P : \mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/N\mathbb{Z}$ be a polynomial of degree d in the field $\mathbb{Z}/N\mathbb{Z}$, and let $f(x) := e_N(P(x))$, thus f is a bounded function. One can easily verify that $\|f\|_{U^{k-1}} = 1$ when $d \leqslant k-2$ (basically because the $(k-1)^{th}$ derivative of P vanishes), so that P is not uniform of any order d or greater. (In fact, one has the more general statement that $\|fg\|_{U^{k-1}} = \|g\|_{U^{k-1}}$ for arbitrary g and whenever $d \leqslant k-2$; thus the U^{k-1} norm is invariant under polynomial phase modulations of degree $k-2$ or less). On the other hand, one can verify that $\|f\|_{U^{k-1}} = O_d(N^{-1/2})$ when $d > k-2$; this is easiest to accomplish when $d = k-1$, and the remaining cases follow by monotonicity (or van der Corput type arguments for Weyl sums). Thus P is uniform of order $d-1$ or less. The intuition to have here is that a bounded function is (heuristically) uniform of order d iff its phase is "orthogonal" to all polynmial phases of degree d or less. In the $d = 1$ case this intuition is precise: linear uniformity corresponds to being orthogonal to linear phase functions, as the estimates* (2.11) *already attest to. When $d \geqslant 2$ however this intuition is harder to pin down, and the theory is still not completely understood.*

Now consider a quadratic polynomial $P(x)$, with corresponding quadratic phase function $f(x) := e_N(P(x))$. From the identity

$$P(x) - 3P(x+r) + 3P(x+2r) - P(x+3r) = 0$$

(which reflects the fact that the third derivative of P), we observe that

$$\Lambda_3(f, \overline{f}, f, \overline{f}) = 1.$$

Thus f is non-negligible for the purposes of computing the Λ_3 form. This is despite f being linearly uniform (all the Fourier coefficients of f is $O(N^{-1/2})$, as one sees from the classical theory of Gauss sums). This shows that for the purposes of analyzing Λ_3, it *is really quadratic uniformity which is the concept to be studied, not linear uniformity. Similarly, the concept of being Gowers uniform of order $k-2$ is the one which is related to the form Λ_k, which in turn counts arithmetic progressions of length k.*

4. Progressions of length 4

With the above machinery, we can now sketch two different proofs of Szeméredi's theorem for progressions of length 4. (These arguments also extend, with some additional difficulties, to higher k, but we will not discuss these technicalities here). The first proof we present is due to Gowers [16] and can be viewed as a generalization of Roth's Fourier-analytic argument, being a density-incrementation argument using quadratic Fourier analysis instead of linear Fourier analysis. The second proof is adapted from that in [41], which in turn is based on the original ergodic theory arguments of Furstenberg and co-authors [10], [11]. It is a generalization of the second proof of Roth's theorem given earlier; in particular, it is is an energy-incrementation argument based on the decomposition of an arbitrary function into a "almost periodic function of order 2" and a quadratically uniform function.

We begin by discussing Gowers' proof, though we shall omit many of the details which pertain to arithmetic combinatorics. Once again, we have a subset A of $[1, N]$, which we embed into a cyclic group $\mathbb{Z}/p\mathbb{Z}$ of prime order. We split $f = 1_A = g+b$, where $g = \mathbb{E}(f)$ and $b = f - \mathbb{E}(f)$. If b is quadratically uniform in the sense that $\|b\|_{U^3}$ is suitably small (less than $c\delta^C$ for some absolute constants $c, C > 0$) then, by using Lemma 3.4 to develop an analogue of Proposition 2.3, then one can easily obtain non-trivial lower bounds for $\Lambda_4(f, f, f, f)$ and thus establish plenty of arithmetic progressions of length 4 in A.

The difficulty comes in the "hard case", when b is not quadratically uniform, so that $\|b\|_{U^3}$ is relatively large. The difficulty here is that unlike the U^2 norm, which is the l^4 norm of the Fourier transform, the U^3 norm is not easily related to the Fourier transform; for instance in Example 3.5 we saw that there were functions which had very small Fourier transform but had large U^3 norm. Nevertheless, it is still possible to use this information to deduce some structural information about A. The situation can be clarified somewhat by considering a model problem, which is to determine all functions of the form $b = e_p(\phi(x))$ which had the maximal U^3 norm of 1, where $\phi : \mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ is a phase function. Expanding out the U^3 norm, we see that this is equivalent to asking that

$$\phi(x+r+s+t)-\phi(x+r+s)-\phi(x+r+t)-\phi(x+s+t)+\phi(x+r)+\phi(x+s)+\phi(x+t)-\phi(x) = 0 \quad (4.1)$$

for all $x, r, s, t \in \mathbb{Z}/p\mathbb{Z}$. This is an "arithmetic" way of asserting that the third derivative of ϕ vanishes. It in fact implies that ϕ is a quadratic polynomial, $\phi(x) = ax^2 + bx + c$ (whereas in contrast, the assertion that b would have a maximal Fourier coefficient of 1 is equivalent to asserting that ϕ is a linear polynomial). To see this, let us adopt the notation that for any function $f : \mathbb{Z}/p\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ and any shift $h \in \mathbb{Z}/p\mathbb{Z}$, that $f_h : \mathbb{Z}/p\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ denotes the "derivative" $f_h(x) := f(x+h) - f(x)$. Then we have

$$\phi_h(x+s+t) - \phi_h(x+s) - \phi_h(x+t) + \phi_h(x) = 0 \text{ for all } x, s, t \in \mathbb{Z}/p\mathbb{Z}. \quad (4.2)$$

It is easy to see that this implies that ϕ_h is linear, i.e. we have

$$\phi_h(x) = a(h)x + c(h) \text{ for all } x, h \in \mathbb{Z}/p\mathbb{Z} \quad (4.3)$$

for some $a(h)$, $c(h) \in \mathbb{Z}/p\mathbb{Z}$. (This is easiest seen by first subtracting $\phi_h(0)$ from ϕ_h, at which point ϕ_h becomes additive). To conclude from this that ϕ is quadratic, one would need to firstly show that $a(h)$ and $c(h)$ have some linearity properties in h, and then "integrate" the equation (4.3) to obtain a quadratic expression for $\phi(x)$.

To attain these goals, we rewrite (4.3) as the functional equation

$$a(h)x = \phi(x+h) - \phi(x) - c(h) \text{ for all } x, h \in \mathbb{Z}/p\mathbb{Z}. \quad (4.4)$$

We can isolate $a(h)$ in this equation by taking suitable "derivatives". For instance, if one replaces x by $x+s$ in the above formula to obtain

$$a(h)(x+s) = \phi(x+s+h) - \phi(x+s) - c(h) \text{ for all } x, h, s \in \mathbb{Z}/p\mathbb{Z}.$$

and then subtracts the two equations, one obtains

$$a(h)s = \phi_s(x+h) - \phi_s(x) \text{ for all } x, h, s \in \mathbb{Z}/p\mathbb{Z} \quad (4.5)$$

thus eliminating the unknown function $c(h)$. Similarly, by replacing h by $h+t$ and then subtracting, we can eliminate the $\phi_s(x)$ term to obtain

$$a_t(h)s = \phi_{st}(x+h) \text{ for all } x, h, s, t \in \mathbb{Z}/p\mathbb{Z}.$$

Finally, by replacing x, h by $x-u, h+u$ and subtracting again to eliminate the $\phi_{st}(x+h)$ term, one obtains

$$a_{tu}(h)s = 0 \text{ for all } h, s, t, u \in \mathbb{Z}/p\mathbb{Z} \quad (4.6)$$

and thus a obeys the functional equation

$$a(h+t+u) - a(h+t) - a(h+u) + a(h) = 0 \text{ for all } h, u, t \in \mathbb{Z}/p\mathbb{Z} \quad (4.7)$$

which as observed earlier implies that $a(h)$ is linear, thus

$$a(h) = \alpha h + \beta \text{ for some } \alpha, \beta \in \mathbb{Z}/p\mathbb{Z}. \quad (4.8)$$

(One can in fact force β to equal zero, basically because $a(0) = 0$, but we will not do so here). Now the function αhx can be explicitly integrated (modulo a lower order term) using the quadratic primitive

$$F(x) := \frac{\alpha}{2}x^2, \quad (4.9)$$

in the sense that $F_h(x) = \alpha hx + \frac{\alpha}{2}h^2$. Thus if we define $\phi'(x) := \phi(x) - F(x)$ and $\phi''(x) := \phi(x) - F(x) + \beta x$, then by (4.3), $\tilde{\phi}$ obeys the functional equation

$$\phi'(x+h) - \phi''(x) = \beta x + c(h) - \frac{\alpha}{2}h^2 \text{ for all } x, h \in \mathbb{Z}/p\mathbb{Z}. \quad (4.10)$$

Replacing x by $x+k$ and subtracting, we obtain that

$$\phi'(x+h+k) - \phi'(x+h) - \phi''(x+k) + \phi''(x) = 0 \text{ for all } x, h, k \in \mathbb{Z}/p\mathbb{Z}$$

which then implies that ϕ' and ϕ'' is linear. Since $\phi = \phi' + F$, we thus see that ϕ is quadratic as claimed.

This concludes the treatment of the model problem. Thanks to the work of Gowers [16], it turns out that the general strategy used to solve this model problem can also be used to handle the general case. Indeed, if a function b has large U^3 norm (where by "large" we mean "larger than $C^{-1}\delta^C$ for some absolute constant $C > 0$"), then by (3.5) the function $b(x+h)\overline{b(x)}$ will have large U^2 norm for a large percentage of $h \in \mathbb{Z}/p\mathbb{Z}$ (this is the analogue of (4.2)). Since U^2 norms imply large Fourier coefficients, we thus see that for all h in a large fraction $H \subset \mathbb{Z}/p\mathbb{Z}$ of $\mathbb{Z}/p\mathbb{Z}$ we can find $a(h), c(h) \in \mathbb{Z}/p\mathbb{Z}$ such that

$$\Re\mathbb{E}(b(x+h)\overline{b(x)}e_p(a(h)x + c(h))|x \in \mathbb{Z}/p\mathbb{Z}) \geqslant C^{-1}\delta^C \quad (4.11)$$

and hence

$$|\mathbb{E}(b(x+h)\overline{b(x)}e_p(-(a(h)x + c(h)))1_H(h)|x, h \in \mathbb{Z}/p\mathbb{Z})| \geqslant C^{-1}\delta^C.$$

As with the model problem, the task would now be to obtain some linearity control on a. This can be obtained by a Cauchy-Schwarz argument; there are a number of permutations of this argument, but we shall give one which is based on the van der Corput lemma, Lemma 3.1. Let us first change variables $x = y_1 - y_2$, $h = y_2 - y_3$ to obtain

$$|\mathbb{E}(b(y_1-y_3)\overline{b(y_1-y_2)}1_H(y_2-y_3)e_p(-c(y_2-y_3))K(y_1, y_2, y_3)|y_1, y_2, y_3 \in \mathbb{Z}/p\mathbb{Z})| \geqslant C^{-1}\delta^C,$$

where

$$K(y_1, y_2, y_3) := e_p(-a(y_2 - y_3)(y_1 - y_2))1_H(y_2 - y_3).$$

If we then apply Lemma 3.1, we conclude that

$$\|K\|_{\Box^3} \geqslant C^{-1}\delta^C.$$

Raising this to the eighth power and expanding out the left-hand side, one eventually obtains (after some change of variables)

$$\mathbb{E}(e_p(-(a(h+t+u)-a(h+t)-a(h+u)+a(h))s)1_{h,h+u,h+t,h+t+u\in H}|h, t, s, u \in \mathbb{Z}/p\mathbb{Z}) \geqslant C^{-1}\delta^C$$

(this is the analogue of (4.6)). The average in s can be computed explicitly, and we then obtain

$$\mathbb{P}(h, h+t, h+u, h+t+u \in H; a(h+t+u)-a(h+t)-a(h+u)+a(h) = 0|h, t, u \in \mathbb{Z}/p\mathbb{Z}) \geqslant C^{-1}\delta^C \quad (4.12)$$

(cf. (4.7)). This is now a purely arithmetic-combinatorial statement about a, involving no oscillation; it says that a behaves like an (affine-)linear function "a significant fraction of the time". In analogy with (4.8) It is then tempting to conjecture from this that $a(h)$ should in fact *equal* an affine linear function $\alpha h + \beta$ for a significant fraction of the time, i.e. we should be able to find $\alpha, \beta \in \mathbb{Z}/p\mathbb{Z}$ such that

$$\mathbb{P}(h \in H; a(h) = \alpha h + \beta|h \in \mathbb{Z}/p\mathbb{Z}) \geqslant C^{-1}\delta^C \quad (4.13)$$

(note that in the converse direction, that one can use (4.13) and a Cauchy-Schwarz argument to obtain (4.12)). Suppose for the moment that one could indeed deduce (4.13) from (4.12). Then we can introduce the primitive function (4.9) as before, and define $b'(x) := b(x)e_p(-F(x))$ and $b''(x) := b'(x)e_p(\beta x)$; we then see from (4.11) that

$$\Re\mathbb{E}(b'(x+h)\overline{b''(x)}e_p(\frac{\alpha}{2}h^2 - c(h))|x \in \mathbb{Z}/p\mathbb{Z}) \geqslant C^{-1}\delta^C$$

for all $h \in H$ (cf. (4.10)). In particular we see that

$$|\mathbb{E}(b'(x+h)\overline{b''(x)}|x \in \mathbb{Z}/p\mathbb{Z})| \geqslant C^{-1}\delta^C.$$

Taking L^2 norms of both sides and using Plancherel, we obtain

$$\|\widehat{b'}\widehat{b''}\|_{l^2} \geqslant C^{-1}\delta^C,$$

and thus by Hölder's inequality

$$\|b'\|_{U^2}^4 \geqslant C^{-1}\delta^C.$$

To summarize, we started with a function b with large U^3 norm, and then were able to locate a quadratic modulation b' of b which in fact had large U^2 norm. Since we already know that a large U^2 norm would imply a large Fourier coefficient, we could thus deduce the existence of a $\xi \in \mathbb{Z}/p\mathbb{Z}$ such that $\widehat{b'}(\xi)$ is large, which would then imply that the original function b had large correlation with a quadratic phase function $\chi(x) := e_p(P(x))$ for some quadratic polynomial $P : \mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$, thus $|\langle b, \chi\rangle| \geqslant C^{-1}\delta^C$. One can now proceed as in the density increment proof of Roth's theorem, but with the Bohr sets in $\mathcal{B}_{\varepsilon,\chi}$ now being replaced by "quadratic Bohr sets" . This eventually gives us a density increment of the form (2.13) on a quadratic Bohr set $\chi^{-1}(Q)$; one can then use Weyl's theorem on equidistribution of quadratic polynomials mod p to locate a reasonably long arithmetic progression (of length at least cN^c for some absolute constant $c > 0$, if N is sufficiently large depending on δ) on which one has a density increment, at which point we may repeat Roth's argument. We omit the details, referring the reader instead to [16].

We return briefly now to a step glossed over in the above sketch, namely the deduction of (4.13) from (4.12). As it turns out, this implication is false as stated; it is possible for a to be additive in the sense of (4.12) without being approximately linear in the sense of (4.13), because a may instead be behaving like a "higher-dimensional" linear function. An example of this is as follows. Let M be an integer between $\sqrt{p}/4$ and $\sqrt{p}/2$, let $H := \{n + 2Mm : 1 \leqslant n, m \leqslant M\}$, and let $a : H \to \mathbb{Z}/p\mathbb{Z}$ be the function $a(n + 2Mm) = \alpha n + \beta m$ for some fixed $\alpha, \beta \in \mathbb{Z}/p\mathbb{Z}$. Then one can easily verify that a obeys the property (4.12) but not (4.13) (if $\beta \neq 2M\alpha$). The set H is an example of a *two-dimensional arithmetic progression*, and the function a given here is a *generalized linear function* on this progression; more generally one can define the notion of a generalized arithmetic progression (of arbitrary dimension), and of a generalized linear function on this progression; it is possible then to obtain a deduction of the form (4.12) $\implies$ (4.13) but with the role of $\alpha h + \beta$ being played by these generalized linear functions; also, for technical reasons (having to do with relatively poor constants in a certain inverse theorem from additive combinatorics known as Freiman's theorem) one must with the lower bound of $C^{-1}\delta^C$ by a smaller quantity such as $\exp(-C\delta^{-C})$; it is not known whether this exponential loss has to be removed. The deduction here requires a combination of techniques from combinatorial graph theory, probabilistic combinatorics, Fourier analysis, and the geometry of lattices and Bohr sets; it is somewhat involved and we will not go into the details here, referring the reader instead to [16].

The remainder of Gowers' argument in [16] is concerned with how to use the fact that a is approximately equal to a higher-dimensional linear function to again deduce a density increment of A on some sub-progression. This is again done mainly by Weyl's theory of uniform distribution; however in [25] an alternate argument was developed, which is based on locating a primitive F to a. This argument closely mimics the one given in the one-dimensional case when $a(h) \approx \alpha h + \beta$; however, there is an additional difficulty in the higher-dimensional case, namely that not every linear function has a primitive; instead, only the "self-adjoint" linear functions do. This has to do with the fact that quadratic forms in higher dimensions (the analogue of quadratic polynomials in one dimension) are associated to symmetric matrices rather than general matrices. Fortunately, one can show that the function a does indeed obey the required symmetry property. Rather than give the precise statement and proof of this assertion in detail, we sketch how it works in a model case. Here we consider solutions to the equation

(4.4), but now x, h take values in a vector space $V := (\mathbb{Z}/p\mathbb{Z})^n$, and $a(h)$ is now a linear transformation from V to $\mathbb{Z}/p\mathbb{Z}$. By arguing as before, we conclude that $a(h) = \alpha h + \beta$, where β is now a linear transformation from V to $\mathbb{Z}/p\mathbb{Z}$, and α is a bilinear form from $V \times V$ to $\mathbb{Z}/p\mathbb{Z}$. Inserting this back into (4.5), we obtain

$$\alpha(h, s) + \beta s = \phi(x + h + s) - \phi(x + h) - \phi(x + s) + \phi(x) \text{ for all } x, h, s \in V.$$

Now we proceed a little differently to before. If we replace h, s by $h + u, s - u$ and subtract, we obtain

$$\alpha(h + u, s - u) - \alpha(h, s) - \beta u = -\phi_u(x + h) - \phi_{-u}(x + s) \text{ for all } x, h, s, u \in V.$$

If now we replace x, h, s by $x + t, h - t, s - t$ and subtract, we obtain

$$\alpha(h+u-t, s-u-t) - \alpha(h-t, u-t) - \alpha(h+u, s-u) + \alpha(h, s) = 0 \text{ for all } h, s, u, t \in V.$$

Using the bilinearity of α, this simplifies to

$$\alpha(t, u) - \alpha(u, t) = 0 \text{ for all } u, t \in V$$

which shows that α is symmetric. In particular this allows us to construct a primitive F by the formula $F(x) := \frac{1}{2}\alpha(x, x)$, and the previous argument now proceeds as before. Back in the original setting of a function b with large U^3 norm, an analogous argument allows us to locate a "generalized quadratic polynomial phase function" $\chi(x) := e_p(P(x))$ such that $\langle b, \chi \rangle$ is somewhat large; see [25] for a rigorous statement and proof of this "inverse theorem for the U^3 norm". (Interestingly, there are some closely related results arising from ergodic theory; see [29], [47]).

This concludes our discussion of Gowers' proof of Szemerédi's theorem for progressions of length 4; the argument also extends to higher k (see [17]) though with some non-trivial additional difficulties; also, it is not at present clear whether the higher U^d norms also enjoy an inverse theorem. We now briefly discuss another proof of this theorem, which extends the energy increment proof for progressions of length three discussed earlier. There are many proofs in this spirit, starting with the work of Furstenberg [10], [11] (and a related energy-incrementation argument also appears in [39]); we shall loosely follow the version of this argument from [41]. For sake of simplicity we shall confine our discussion to the $k = 4$ case only.

As it turns out, large portions of the energy increment proof generalize without difficulty to obtain progressions of arbitrary length. The main difficulty is to replace the concept of an (δ, K)-almost periodic function with a "higher order" generalization. The definition given in Definition 2.6 relies too heavily on linear phase functions, and we have already seen some difficulties in extending that concept to higher orders; for instance, we still do not have a satisfactory theory of what a "quadratically quasiperiodic function" should be, although there are some very promising developments in the ergodic theory of nilfactors (see e.g. [29], [47], [48]) which should shed light on this question very soon. However, it is well understood by now how to generalize the more general concept of an almost periodic function. In ergodic theory, a function f in a measure-preserving system $(X, \mathcal{B}, \mu, T)$ is said to be almost periodic if the orbit $\{T^n f : n \in \mathbb{Z}\}$ is precompact, and in particular can be approximated to arbitrary accuracy by a subset of a finite-dimensional space. In the discrete setting of $\mathbb{Z}/N\mathbb{Z}$, every function is periodic of order N and is thus, technically speaking, every function is almost periodic. However one can still extract a useful concept of almost periodicity by making the concept of "precompact" more quantitative. One such way of doing so is

Definition 4.1 (Uniform almost periodicity norms)**.** [41] If A is a shift-invariant Banach algebra of functions on $\mathbb{Z}_N$, we define the space $UAP[A]$ to be the space of all functions F for which the orbit $\{T^n F : n \in \mathbb{Z}\}$ has a representation of the form

$$T^n F = M\mathbb{E}(c_{n,h} g_h) \text{ for all } n \in \mathbb{Z}_N \quad (4.14)$$

where $M \geqslant 0$, H is a finite non-empty set, $g = (g_h)_{h\in H}$ is a collection of bounded functions, $c = (c_{n,h})_{n\in\mathbb{Z}_N, h\in H}$ is a collection of functions in A with $\|c_{n,h}\|_A \leqslant 1$, and h is a random variable taking values in H. We define the norm $\|F\|_{UAP[A]}$ to be the infimum of M over all possible representations of this form.

The formula (4.14) is a quantitative assertion that the orbit $\{T^n F : n \in \mathbb{Z}\}$ can be represented efficiently by what is essentially a finite-dimensional approxmation, and is thus an assertion of precompactness "relative to A". It can be shown (see [41]) that $UAP[A]$ is a shift-invariant Banach algebra. If we let A be the trivial Banach algebra of constant functions (so that the $c_{n,h}$ are constants, with $\|c_{n,h}\|_A = |c_{n,h}|$) then we abbreviate $UAP[A]$ as UAP^1, and refer to functions with bounded UAP^1 norm as *linearly uniformly almost periodic*. For instance, one can show that any K-quasiperiodic function is linearly uniformly almost periodic, with a UAP^1 norm of at most K. In particular, linear phase functions are linearly uniformly almost periodic, with a UAP^1 norm of exactly 1.

One can then define the space $UAP^2 := UAP[UAP^1]$ of quadratically uniformly functions, which are roughly speaking the space of functions which are almost periodic relative to the linearly almost periodic functions. For example, consider the function $f(x) := e_N(x^2)$. This function is very far from being linearly almost periodic - in the sense that the UAP^1 norm is huge - because the translates $T^n f(x) = e_N(x^2 + 2nx + n^2)$ are all quite distinct and cannot efficiently be expressed as linear combinations of a small number of functions. On the other hand, we may write $T^n f = c_n g$ where $g := f$ and $c_n(x) = e_N(2nx + n^2)$, and note that each c_n, being a linear phase function, lies in UAP^1 with small norm. Thus this function is quadratically almost periodic; in fact, it lies in UAP^2 with norm 1. The property of being quadratically almost periodic strictly generalizes the concept of a *quadratic eigenfunction* in ergodic theory; see e.g. [47], [48] for further discussion.

The concept of quadratic almost periodicity (bounded UAP^2 norm) is in many ways dual to that of quadratic uniformity (small U^3 norm). We present three results supporting this claim. The first is the duality inequality

$$|\langle f, F\rangle| \leqslant \|f\|_{U^3}\|F\|_{UAP^2},$$

which can be proven by a simple Cauchy-Schwarz argument, see [41]. Secondly, if f is such that $\|f\|_{U^3}, \|f\|_{L^\infty} \leqslant 1$, and we let $\mathcal{D}f$ denote the *dual function*

$$\mathcal{D}f(x) := \mathbb{E}(\overline{f(x+a)f(x+b)f(x+c)}f(x+a+b)f(x+a+c)f(x+b+c)\overline{f(x+a+b+c)}|a,b,c\in\mathbb{Z}_N)$$

then $\mathcal{D}f$ lies in UAP^2 with a norm of at most 1; again, see [41]. Furthermore, we have the correlation identity

$$\langle f, \mathcal{D}f\rangle = \|f\|_{U^3}^8.$$

By using these dual function to replace the role of linear (or quadratic) phase functions, one can obtain the following variant of Proposition 2.11:

Proposition 4.2 (Quantitative Koopman-von Neumann theorem)**.** [41] *Let* $F : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ *be an arbitrary function, let* $0 < \sigma < \delta \leqslant 1$*, and let* f *be any bounded* *non-negative function on* $\mathbb{Z}/N\mathbb{Z}$ *obeying* (2.1). *Then there exists a quantity* $0 < K \leqslant C(F, \delta, \sigma)$ *and a decomposition* $f = g + b$, *where* g *is bounded, non-negative, has mean* $\mathbb{E}(g) = \mathbb{E}(f)$, *and we have the bound*

$$\|b\|_{U^3} \leqslant F(\delta, \sigma, K).$$

Furthermore we have an additional decomposition $g = \tilde{g} + e$ *with* $\tilde{g}$ *non-negative and the bounds*

$$\|\tilde{g}\|_{UAP^2} \leqslant K; \quad \|e\|_{L^2} \leqslant \sigma.$$

The proof of this Proposition proceeds by an energy incrementation argument very similar to Proposition 2.11; one begins with the trivial splitting $f = \mathbb{E}(f) + (f - \mathbb{E}(f))$, and whenever the bad function b fails to be quadratically uniform, one uses the dual function $\mathcal{D}b$ (which is quadratically almost periodic) to refine the σ-algebra used to construct the good function g, thus increasing the energy of g by a non-trivial amount.

By combining this with the generalized von Neumann theorem in Lemma 3.4, we can conclude the proof of Szemerédi's theorem in this case once we show the analogue of Lemma 2.7:

Theorem 4.3 (Almost periodic functions are recurrent)**.** *Let* $g, \tilde{g}$ *be non-negative bounded functions such that we have the estimates*

$$\|g - \tilde{g}\|_{L^2} \leqslant \frac{\delta^2}{4096} \tag{4.15}$$

$$\mathbb{E}(g|\mathbb{Z}_N) \geqslant \delta \tag{4.16}$$

$$\|\tilde{g}\|_{UAP^2} < M \tag{4.17}$$

for some $0 < \delta, M < \infty$. *Then we have*

$$\mathbb{E}(g(x)T^n g(x)T^{2n} g(x)T^{3n} g(x)|x, r \in \mathbb{Z}_N) \geqslant c(\delta, M) \tag{4.18}$$

for some $c_0(\delta, M) > 0$.

The proof of this theorem is the most difficult component of the argument; it uses the uniform almost periodicity control on $\tilde{g}$ to "color" the orbit of $T^n \tilde{g}$ and hence $T^n g$, and then invokes the van der Waerden theorem [44] to extract arithmetic progressions from g. As such, this part of the argument can be considered to be more combinatorial than ergodic or analytic in nature.

5. Progressions in the primes

There are many questions concerning the distribution of the prime numbers (and of various configurations of prime numbers), which has motivated a large portion of analytic number theory. One of the basic results in the subject is of course the *prime number theorem*, which asserts that the number of primes between 1 and N asymptotically approaches $N/\log N$ as $N \to \infty$, or in other words

$$\#\{1 \leqslant n \leqslant N : n \text{ is prime}\} = \frac{N}{\log N}(1 + o(1)),$$

where we use $o(1)$ to denote a quantity which goes to zero as $N \to \infty$.

It is convenient to normalize the prime number theorem in a different form. Define the *von Mangoldt function* $\Lambda : \mathbb{Z}^+ \to \mathbb{R}$ by setting $\Lambda(n) := \log p$ whenever $n = p^j$ is a power of a prime p for some $j \geqslant 1$, and $\Lambda(n) = 0$ otherwise; the significance of this function to number theory lies in the identity

$$\log n = \sum_{d|n} \Lambda(d) \tag{5.1}$$

for all integers n (where the sum is over all integers d dividing n), which is a restatement of the unique factorization theorem. The Von Mangoldt function is essentially supported on the primes (there are also the squares and higher powers of primes, but they are extremely sparse, and in practice are completely negligible, contributing only to the $o(1)$ error terms). Then the prime number theorem is easily seen to be equivalent to

$$\frac{1}{N} \sum_{1 \leqslant n \leqslant N} \Lambda(n) = 1 + o(1).$$

The expression on the left-hand side can be viewed as an average or *expectation* for Λ; we shall emphasize this probabilistic (or ergodic) perspective by writing it as $\mathbb{E}(\Lambda(n) : 1 \leqslant n \leqslant N)$; more generally, we write $\mathbb{E}(f(n)|n \in A)$ for $\frac{1}{|A|} \sum_{n \in A} f(n)$ whenever A is a finite set. Thus Λ has an average value of $1 + o(1)$. The error can be improved; for instance the famous *Riemann hypothesis* is equivalent to the claim

$$\mathbb{E}(\Lambda(n)|1 \leqslant n \leqslant N) = 1 + O(N^{-1/2} \log^2 N).$$

However the improved error estimates are not central to the results we shall discuss here, which are in some sense more focused on the main term in such estimates involving the primes.

Now we consider how to count other patterns inside the primes. One of the oldest (and still unsolved) problems in the field is the *twin prime conjecture*, which asks whether there are an infinite number of primes p such that $p + 2$ is also prime. This would be implied by the statement

$$\liminf_{N \to \infty} \mathbb{E}(\Lambda(n)\Lambda(n + 2) : 1 \leqslant n \leqslant N) > 0.$$

is non-zero for infinitely many N. In fact Hardy and Littlewood made the stronger conjecture, the *Hardy-Littlewood prime tuple conjecture* [26], which would imply the twin prime conjecture, and would indeed verify the stronger estimate

$$\mathbb{E}(\Lambda_N(n)\Lambda_N(n + 2) : 1 \leqslant n \leqslant N) = B_2 + o(1)$$

where B_2 is the *Twin prime constant*

$$\begin{aligned} B_2 &:= \prod_p \frac{\mathbb{P}(n, n+2 \text{ coprime to } p | n \in \mathbb{Z}/p\mathbb{Z})}{\mathbb{P}(n \text{ coprime to } p | n \in \mathbb{Z}/p\mathbb{Z}) \mathbb{P}(n+2 \text{ coprime to } p | n \in \mathbb{Z}/p\mathbb{Z})} \\ &= 2 \prod_{p > 3} \frac{p(p-2)}{(p-1)^2} \\ &= 1.32032\ldots \end{aligned}$$

A related problem is the *strong Goldbach conjecture* - whether every even number (larger than 4) can be written as the sum of two primes; this is essentially the same as asking whether

$$\mathbb{E}(\Lambda(n_1)\Lambda(n_2) : 1 \leqslant n_1, n_2 \leqslant N; n_1 + n_2 = N)$$

is non-zero for all even integers N. The Hardy-Littlewood prime tuple conjecture here would imply that

$$\mathbb{E}(\Lambda(n_1)\Lambda(n_2) : 1 \leqslant n_1, n_2 \leqslant N; n_1 + n_2 = N) = G_2(N) + o(1)$$

where

$$G_2(N) := \prod_p \frac{\mathbb{P}(n_1, n_2 \text{ coprime to } p | n_1, n_2 \in \mathbb{Z}/p\mathbb{Z}; n_1 + n_2 = N)}{\prod_{j=1}^2 \mathbb{P}(n_j \text{ coprime to } p | n_1, n_2 \in \mathbb{Z}/p\mathbb{Z}; n_1 + n_2 = N)}$$

which vanishes when N is odd, and is equal to

$$G_2(N) = B_2 \prod_{p|N; p \geqslant 3} \frac{p-1}{p-2} \geqslant B_2 > 0$$

when N is even. Thus the prime tuple conjecture would imply the strong Goldbach conjecture for sufficiently large N.

The *weak Goldbach conjecture*, which is essentially proven (thanks primarily to the work of Vinogradov [46]), asserts that every odd number N larger than 5 can be written as the sum of three primes. (By "essentially proven" I mean that this conjecture has been verified for $N \leqslant 10^{17}$ and also rigourously proven for $N \geqslant 10^{43000}$). This is essentially asking for the quantity

$$\mathbb{E}(\Lambda(n_1)\Lambda(n_2)\Lambda(n_3) : 1 \leqslant n_1, n_2, n_3 \leqslant N; n_1 + n_2 + n_3 = N)$$

to be positive for all odd integers N. The work of Vinogradov implies

$$\mathbb{E}(\Lambda(n_1)\Lambda(n_2)\Lambda(n_3) : 1 \leqslant n_1, n_2, n_3 \leqslant N; n_1 + n_2 + n_3 = N) = G_3(N) + o(1)$$

where

$$G_3(N) := \prod_p \frac{\mathbb{P}(n_1, n_2, n_3 \text{ coprime to } p | n_1, n_2, n_3 \in \mathbb{Z}/p\mathbb{Z}; n_1 + n_2 + n_3 = N)}{\prod_{j=1}^3 \mathbb{P}(n_j \text{ coprime to } p | n_1, n_2, n_3 \in \mathbb{Z}/p\mathbb{Z}; n_1 + n_2 + n_3 = N)}.$$

This quantity is positive and bounded away from zero for all odd N; thus Vinogradov's work implies the weak Goldbach conjecture for all sufficiently large N; to resolve the remaining cases it is thus natural to try to sharpen the $o(1)$ error term. (For instance, the weak Goldbach conjecture is known to be true if one assumes the generalized Riemann hypothesis, which is extremely useful in improving these error terms). One can generalize Vinogradov's result to sums of k primes for any $k \geqslant 3$; but as we shall explain later, the $k = 2$ case is much more difficult and well beyond the reach of existing techniques.

Now we turn to arithmetic progressions in the primes. In 1933 van der Corput [43] (see also [7]) established that the primes contain infinitely many arithmetic progressions of length 3; indeed we know the significantly stronger statement that the Hardy-Littlewood conjecture holds in this case, or more explicitly that

$$\mathbb{E}(\Lambda(n)\Lambda(n+r)\Lambda(n+2r) : 1 \leqslant n, r \leqslant N) = C_3 + o(1) \quad (5.2)$$

where

$$\begin{aligned} C_3 &:= \prod_p \frac{\mathbb{P}(n, n+r, n+2r \text{ coprime to } p | n, r \in \mathbb{Z}/p\mathbb{Z})}{\prod_{j=0}^2 \mathbb{P}(n+jr \text{ coprime to } p | n, r \in \mathbb{Z}/p\mathbb{Z})} \\ &= \frac{3}{2} \prod_{p \geqslant 5} (1 + \frac{1}{(p-1)^3}) \\ &= 1.534\ldots \end{aligned}$$

More generally, the Hardy-Littlewood prime tuple conjecture implies that

$$\mathbb{E}(\Lambda(n) \ldots \Lambda(n+(k-1)r) : 1 \leqslant n, r \leqslant N) = C_k + o_k(1) \tag{5.3}$$

for all $k \geqslant 0$ (with the error term $o_k(1)$ depending on k), where C_k is the constant

$$C_k = \prod_p \frac{\mathbb{P}(n, \ldots, n+(k-1)r \text{ coprime to } p | n, r \in \mathbb{Z}/p\mathbb{Z})}{\prod_{j=0}^{k-1} \mathbb{P}(n+jr \text{ coprime to } p | n, r \in \mathbb{Z}/p\mathbb{Z})}$$

which is explicitly computable for each k. The case $k = 0$ is trivial, the cases $k = 1, 2$ follow from the prime number theorem, and the case $k = 3$ is just (5.2). More recently, we have the following results:

Theorem 5.1. [23], [25] *The conjecture* (5.3) *is also true for* $k = 4$ *(so there are infinitely many prime arithmetic progressions of length 4). Furthermore, for all* $k \geqslant 0$ *we have*

$$\mathbb{E}(\Lambda(n) \ldots \Lambda(n+(k-1)r) : 1 \leqslant n, r \leqslant N) > c_k + o_k(1) \tag{5.4}$$

for some explicit constant $c_k > 0$ *(which is unfortunately much smaller than* C_k*). This weaker statement still suffices to establish infinitely many prime arithmetic progressions of length* k*.*

All of these results have the flavor of "establish bounds or asymptotics for multilinear averages of Λ". However, some are significantly harder than others, depending on the exact structure of the multilinear average involved. As mentioned earlier, the situation has some parallels with the linear, bilinear, and trilinear Hilbert transform in harmonic analysis; while these expressions are formally very similar in structure, the analytical treatment of each one in the sequence has proven to be significantly harder than the previous one, for instance no L^p estimates for the trilinear Hilbert transform are currently known. A certain subclass of these multilinear averages (the "rank one" averages involving three or more copies of Λ) can be treated by Fourier methods; this includes Vinogradov's theorem and van der Corput's theorem, and see also [2] for further discussion. However, it is by now well established that these techniques cannot directly extend to handle other multilinear averages. The $k = 4$ result in Theorem 5.1 requires a "quadratic" generalization of Fourier analysis, pioneered by Gowers [16], but still in a very early stage of development. The higher cases $k \geqslant 5$ could in principle be treated by polynomial Fourier analysis, of the type developed in [17]; this would likely establish (5.3) for all k, this project is currently a work in progress with the author and Ben Green. Instead, we use an alternate argument based on ergodic theory which is technically simpler but only gives the weaker result (5.4).

There are two main strategies to obtain progressions:

- (Uniformity strategy) Attempt to approximate Λ by some averaged version $\mathbb{E}(\Lambda|\mathcal{B})$ of itself, in such a manner that $\Lambda - \mathbb{E}(\Lambda|\mathcal{B})$ is uniform of the correct order (linearly uniform for $k = 3$, quadratically uniform for $k = 4$). This requires one to estimate exponential sums such as $\sum_n \Lambda(n)e(n\theta)$ or $\sum_n \Lambda(n)e(P(n))$ where P is a polynomial or "generalized polynomial).

- (Szemerédi strategy) Attempt to leverage Szemerédi's theorem (or in the case of progressions of length three, Roth's theorem) in order to obtain arithmetic progressions regardless of whether Λ is uniform or not.

In the case of progressions of length three, the uniformity strategy (more commonly known in this context as the *Hardy-Littlewood circle method*) was developed far earlier than the Szemerédi strategy. It gives sharper results (in particular, it yields the asymptotic (5.3)), but is technically more difficult to implement. We now briefly discuss each of these strategies in turn.

6. The uniformity strategy

We begin by discussing the uniformity strategy. We shall eschew the traditional framework of the Hardy-Littlewood circle method (which is only effective for the $k = 3$ case) and present this strategy in a language which more easily lends itself to generalization to higher k.

The circle method relies on Fourier analysis on the integers $\mathbb{Z}$ (so that the dual group is the unit circle S^1, hence the terminology "circle method"). For us it will be slightly more convenient to work in the cyclic group $\mathbb{Z}/N\mathbb{Z}$, which is self-dual. To simplify the exposition we shall pretend that Λ is actually a function on $\mathbb{Z}/N\mathbb{Z}$ rather than $\mathbb{Z}$. In practice one would have to justify this by a truncation trick, for instance cutting off Λ to $\{1, \ldots, N/3\}$ (possibly using a smooth cutoff function) and then transferring this to $\mathbb{Z}/N\mathbb{Z}$; this type of "transference" is quite standard and introduces no substantial difficulties, and so we shall gloss over this entire issue.

Using the above "cheat", we can morally rewrite (5.3) as

$$\mathbb{E}(\Lambda(n)\ldots\Lambda(n+(k-1)r) : n, r \in \mathbb{Z}/N\mathbb{Z}) = C_k + o_k(1).$$

Let us first discuss the $k = 3$ case (i.e. (5.2)), which with our new cheat becomes

$$\mathbb{E}(\Lambda(n)\Lambda(n+r)\Lambda(n+2r) : n, r \in \mathbb{Z}/N\mathbb{Z}) = C_3 + o(1).$$

. The strategy is to use some variant[7] of Proposition 2.3. More specifically, we would seek to approximate Λ by an averaged version $\mathbb{E}(\Lambda|\mathcal{B})$ such that we have a uniformity estimate

$$\|(\Lambda - \mathbb{E}(\Lambda|\mathcal{B}))^\wedge\|_\infty = o(1), \tag{6.1}$$

which (by a suitable variant of Proposition 2.3) should imply

$$\begin{aligned}\mathbb{E}(\Lambda(n)\Lambda(n+r)\Lambda(n+2r) &: n, r \in \mathbb{Z}/N\mathbb{Z})\\ &= \mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n)\mathbb{E}(\Lambda|\mathcal{B})(n+r)\mathbb{E}(\Lambda|\mathcal{B})(n+2r) : n, r \in \mathbb{Z}/N\mathbb{Z}) + o(1)\end{aligned}$$

and then one only has to prove (5.2) for the averaged function $\mathbb{E}(\Lambda|\mathcal{B})$:

$$\mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n)\mathbb{E}(\Lambda|\mathcal{B})(n+r)\mathbb{E}(\Lambda|\mathcal{B})(n+2r) : n, r \in \mathbb{Z}/N\mathbb{Z}) = C_3 + o(1). \tag{6.2}$$

[7]Strictly speaking, one has to replace this Proposition by a weighted variant to cope with the fact that Λ is not a bounded function. This can be done by using a suitable weight function Λ_R which is adapted to "almost primes", and which among other things obeys a good Fourier restriction theorem which allows one to transfer Proposition 2.3 to the weighted setting. See [20], [24] for further discussion of this issue.

The first issue is to decide what function $\mathbb{E}(\Lambda|\mathcal{B})$ to use as the approximant to Λ. In order to establish (6.2) we would like $\mathbb{E}(\Lambda|\mathcal{B})$ to have low "complexity" - in particular, it should be far more regular than Λ itself - but not so simple that the approximation to Λ is poor in the sense that (6.1) fails.

Let us understand what (6.1) means. We can rewrite it as

$$\widehat{\mathbb{E}(\Lambda|\mathcal{B})}(\xi) = \hat{\Lambda}(\xi) + o(1) \text{ for all } \xi \in \mathbb{Z}/N\mathbb{Z}$$

or in other words

$$\mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n)e_N(-n\xi)|n \in \mathbb{Z}/N\mathbb{Z}) = \mathbb{E}(\Lambda(n)e_N(-n\xi)|n \in \mathbb{Z}/N\mathbb{Z}) + o(1) \text{ for all } \xi \in \mathbb{Z}/N\mathbb{Z}. \tag{6.3}$$

This gives us some clues as to what kind of approximation $\mathbb{E}(\Lambda|\mathcal{B})$ we should choose. For instance, setting $\xi = 0$ and using the prime number theorem $\mathbb{E}(\Lambda(n)|n \in \mathbb{Z}/N\mathbb{Z}) = 1 + o(1)$, we see that we need $\mathbb{E}(\Lambda|\mathcal{B})$ to obey the condition

$$\mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n)|n \in \mathbb{Z}/N\mathbb{Z}) = 1 + o(1).$$

This suggests using the constant function 1 (or perhaps $\mathbb{E}(\Lambda) = 1 + o(1)$) as the approximating function $\mathbb{E}(\Lambda|\mathcal{B})$; this corresponds to interpreting $\mathcal{B}$ as the trivial σ-algebra $\mathcal{B}_1 := \{\emptyset, \mathbb{Z}/N\mathbb{Z}\}$. For this approximation, the left-hand side of (6.2) is very easy to compute, indeed it is just $1+o(1)$. Unfortunately, while (6.3) is true for this approximation when $\xi = 0$, it is not true for some other values of ξ. Take for instance $\xi = \lfloor N/2 \rfloor$. Then $e_N(-n\xi)$ is essentially $+1$ when n is even and -1 when n is odd, and so if $\mathbb{E}(\Lambda|\mathcal{B}_1)$ were constant then the left-hand side of (6.3) would vanish. On the other hand, the right-hand side of (6.3) is large and negative, because Λ is overwhelmingly supported on the odd numbers rather than the even numbers. Thus we must modify the approximant $\mathbb{E}(\Lambda|\mathcal{B}_1)$ to reflect this "bias" that Λ has towards being odd. The easiest way to fix this is to refine the σ-algebra $\mathcal{B}_1$ to include the odd and even numbers. In other words, if we now let $\mathcal{B}_2$ be the σ-algebra generated by $\mathcal{B}_1$ and the residue classes mod 2 (i.e. the odd and even numbers), then we can use $\mathbb{E}(\Lambda|\mathcal{B}_2)$ as our approximant. By the prime number theorem (and the fact that almost all primes are odd), we know that this function is $2 + o(1)$ on the odd numbers and $o(1)$ on the even numbers. One can now also check that (6.3) is now true when ξ is close to zero or close to $N/2$. Furthermore, the left-hand side of (6.2) is quite easy to compute, it is

$$\frac{\mathbb{P}(n, n+r, n+2r \text{ coprime to } 2|n, r \in \mathbb{Z}/2\mathbb{Z})}{\prod_{j=0}^{2} \mathbb{P}(n+jr \text{ coprime to } 2|n, r \in \mathbb{Z}/2\mathbb{Z})} + o(1) = 2 + o(1).$$

Unfortunately, there are still some further Fourier-analytic biases in Λ which are not detected by the approximation $\mathbb{E}(\Lambda|\mathcal{B}_2)$, for instance the fact that Λ is concentrated in the residue classes $1 \mod 3$ and $2 \mod 3$ and nearly vanishes on the residue class 0 mod 3 will cause the Fourier coefficients of Λ to be rather large for ξ near $N/3$ and $2N/3$, whereas $\mathbb{E}(\Lambda|\mathcal{B}_2)$ is uniformly distributed among all three residue classes $\mod 3$ and thus has a negligible Fourier coefficient at those frequencies. One can address this failure of (6.3) by refining the approximation $\mathbb{E}(\Lambda|\mathcal{B}_2)$ further to $\mathbb{E}(\Lambda|\mathcal{B}_3)$, where $\mathcal{B}_3$ is the σ-algebra formed by adjoining the residue classes modulo 3 to $\mathcal{B}_2$ (or in other words, $\mathcal{B}_3$ is the σ-algebra generated by the residue classes modulo 6). Then one can show that (6.3) now holds for all ξ near multiples of $N/6$. Furthermore, one has $\mathbb{E}(\Lambda|\mathcal{B}_3)(n) = 3 + o(1)$ when n is coprime to 6 and $\mathbb{E}(\Lambda|\mathcal{B}_3) = o(1)$ otherwise; this follows from the prime number theorem combined with *Dirichlet's theorem*, which asserts that Λ is uniformly distributed among those residue classes modulo m which are coprime to m, as long as N is sufficiently large compared to m (here we take $m = 6$). Because of this, one can compute (using the Chinese remainder theorem) that the left-hand side of (6.2) is now

$$\prod_{p=2,3} \frac{\mathbb{P}(n, n+r, n+2r \text{ coprime to } p | n, r \in \mathbb{Z}/2\mathbb{Z})}{\prod_{j=0}^{2} \mathbb{P}(n+jr \text{ coprime to } p | n, r \in \mathbb{Z}/2\mathbb{Z})} + o(1) = \frac{3}{2} + o(1).$$

One can of course continue in this fashion. Let $w = w(N)$ be a slowly growing function of N, e.g. $w = \log\log N$, and let W be the product of all the primes less than w. We let $\mathcal{B}_w$ be the σ-algebra formed by the residue classes modulo W, then we use $\mathbb{E}(\Lambda|\mathcal{B}_w)$ as our approximant. From Dirichlet's theorem, one can show (if w is sufficiently slowly growing in N) that $\mathbb{E}(\Lambda|\mathcal{B}_w)(n) = \frac{W}{\phi(W)} + o(1)$ if n is coprime to W, and $\mathbb{E}(\Lambda|\mathcal{B}_w)(n) = o(1)$ otherwise; here $\phi(W)$ is the Euler totient function of W, i.e. the number of integers in $\{1, \ldots, W\}$ which are coprime to W. From the Chinese remainder theorem, the left-hand side of (6.2) can be computed as

$$\prod_{p \leqslant w} \frac{\mathbb{P}(n, n+r, n+2r \text{ coprime to } p | n, r \in \mathbb{Z}/2\mathbb{Z})}{\prod_{j=0}^{2} \mathbb{P}(n+jr \text{ coprime to } p | n, r \in \mathbb{Z}/2\mathbb{Z})} + o(1) = C_3 + o(1)$$

since the product is convergent and w tends (slowly) to infinity. Thus it only remains to demonstrate (6.3). This would be easy if w was extremely large (e.g. if $w = \sqrt{N}$, then the sieve of Eratosthenes essentially ensures that $\Lambda = \mathbb{E}(\Lambda|\mathcal{B}_w)$, but unfortunately the error terms blow up long before w reaches this level. Nevertheless, this "W-trick" of removing all the structure from Λ associated to those primes less than w does make the task of (6.3) much easier. Essentially, it means that (6.3) is automatically true whenever ξ is a "major arc frequency", which roughly means that $\xi \approx aN/q$ for some integers a, q with $q \leqslant w$. It thus remains to prove (6.3) when ξ is a "minor arc" frequency, which roughly means that $q\xi$ is not close to zero modulo N for any $q \leqslant w$. In such a case, the left-hand side of (6.3) is very small (by construction of $\mathbb{E}(\Lambda|\mathcal{B}_w)$, and one is reduced to establishing enough cancellation in the sum $\sum_{n<N} \Lambda(n) e_N(-n\xi)$ to ensure that it is $o(N)$. (Note that the trivial bound coming from using absolute values and the prime number theorem is $O(N)$).

To do this, one must finally use some deeper structure of the function $\Lambda(n)$, beyond the prime number theorem and Dirichlet's theorem. This was first done by Vinogradov, with later simplifications by Vaughan and other authors; we present a vastly oversimplified sketch of the main idea here. The starting point is the identity (5.1). Solving for n we obtain the formula

$$\Lambda(n) = \sum_{c,d:cd=n} \log c \mu(d),$$

where $\mu(d)$ is the *Möbius function*, defined as $(-1)^m$ if d is the product of m distinct primes, and equal to 0 otherwise. Thus we can write

$$\sum_{n<N} \Lambda(n) e_N(-n\xi) = \sum_{c,d:cd<N} \log c \mu(d) e_N(-cd\xi).$$

The idea is now to view this as a bilinear form acting on the functions log and μ, given by the matrix coefficients $e_N(-cd\xi)$. The hypothesis that ξ is not "minor arc" leads to some almost orthogonality in this matrix (which can be made explicit by the TT^* method), which after some care can eventually lead to the $o(1)$ gain. (This is an oversimplification because the portions of this expression when c or d is small require some additional attention, including a quantitative version of Dirichlet's theorem known as the Siegel-Walfisz theorem; we will not discuss these rather lengthy issues here). This can eventually be used to establish Van der Corput's theorem (5.2).

It turns out that the same ideas can also be pushed (with several additional difficulties) to give the $k = 4$ case of (5.3); it is not yet known whether the arguments can be pushed to general k. By using a result similar to Lemma 3.4, as a substitute for Proposition 2.3, it suffices to find an approximation $\mathbb{E}(\Lambda|\mathcal{B})$ for Λ such that

$$\mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n)\mathbb{E}(\Lambda|\mathcal{B})(n+r)\mathbb{E}(\Lambda|\mathcal{B})(n+2r)\mathbb{E}(\Lambda|\mathcal{B})(n+3r) : n, r \in \mathbb{Z}/N\mathbb{Z}) = C_4 + o(1) \tag{6.4}$$

and

$$\|\Lambda - \mathbb{E}(\Lambda|\mathcal{B})\|_{U^3} = o(1). \tag{6.5}$$

As in the $k = 3$ case, we again invoke the "W-trick" and set $\mathcal{B} = \mathcal{B}_w$ where w is again a slowly growing function of N. When one does so, (6.4) is easy to establish, but (6.5) is still quite difficult. Expanding out the U^3 norm directly gives rise to expressions which are about as complicated to estimate as the original expression in (5.3). However, one can proceed instead by using the inverse theory used in Gowers' proof of Szemerédi's theorem for progressions of length 4. The idea is to assume that $\|\Lambda - \mathbb{E}(\Lambda|\mathcal{B})\|_{U^3}$ is large, say larger than some $\delta > 0$, and arrive at a contradiction. One can repeat the analysis in Gowers' arguments (though one has to introduce weights to deal with the fact that Λ is not bounded) to eventually conclude that

$$\mathbb{E}((\Lambda(n) - \mathbb{E}(\Lambda|\mathcal{B})(n))e_N(P(n))|n \in \mathbb{Z}/N\mathbb{Z}) \geqslant c(\delta) > 0$$

for some "generalized quadratic phase function" $P(n)$; we shall gloss over exactly what "generalized quadratic phase function" means here but one should think of P as being like a quadratic polynomial. Thus to conclude the proof, one needs to extend the linear uniformity estimate (6.3) to the claim that

$$\mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n)e_N(P(n))|n \in \mathbb{Z}/N\mathbb{Z}) = \mathbb{E}(\Lambda(x)e_N(P(n))|n \in \mathbb{Z}/N\mathbb{Z}) + o(1)$$

for all generalized quadratic phase functions P. It turns out that once again one can divide into the case when P is "major arc" - all the non-constant coefficients of P are essentially rational multiples of N with small denominator, and when P is "minor arc" - when at least one of the coefficients behaves "irrationally". The major arc case is again easy, while the minor arc case turns out to be again amenable to the methods of Vinogradov and Vaughan. Here the point is to establish some orthogonality in the matrix coefficients $e_N(P(cd))$. See [25] for further details.

7. The Szemerédi strategy

In principle, the uniformity strategy discussed above should in fact prove (5.3) for all k. However, at present we are restricted to $k \leqslant 4$ because the inverse theorem that passes from large U^{k-1} norm to correlation with a generalized polynomial phase function of order $k-2$ has only been rigorously proven for $k \leqslant 4$. (The analysis in [17] strongly suggests that this inverse theorem should in fact extend to higher k; this is a current work in progress with the author and Ben Green). In particular, while it is conjectured that we in fact have

$$\|\Lambda - \mathbb{E}(\Lambda|\mathcal{B}_w)\|_{U^{k-1}} = o_k(1) \tag{7.1}$$

for all k (which would certainly imply (5.3)), this estimate has not yet been rigorously established.

Nevertheless, one can still achieve the weaker statement (5.4) by using ergodic theory arguments to locate another σ-algebra $\mathcal{B}$ (which could be somewhat finer than $\mathcal{B}_w$) for which the analogue of (7.1) holds. To finish the proof of (5.4), it then remains to show that

$$\mathbb{E}(\mathbb{E}(\Lambda|\mathcal{B})(n) \ldots \mathbb{E}(\Lambda|\mathcal{B})(n+(k-1)r) : 1 \leqslant n, r \leqslant N) > c_k + o_k(1). \tag{7.2}$$

Unfortunately, the structure of the algebra $\mathcal{B}$ is much less well understood than $\mathcal{B}_w$, and as such the function $\mathbb{E}(\Lambda|\mathcal{B})$ is also not very well understood. However, being a conditional expectation of Λ, it is still non-negative, has the same mean (i.e. $1 + o(1)$) as Λ. Crucially, one can also establish that $\mathbb{E}(\Lambda|\mathcal{B})$ is also *bounded* by $O(1)$. By the third version of Szemerédi's theorem, these three facts imply (7.2).

A prototype of this argument is the proof of Theorem 1.4 in [20], which used Fourier analytic methods (but with ergodic ideas lurking under the surface), and as such was limited to the $k = 3$ case. This argument was then simplified and extended in [24]; simultaneously, in [23] the Fourier-analytic components were replaced with ergodic theory arguments which could then extend to general k. Here we shall begin by discussing the general ergodic theory argument, and return to briefly discuss the earlier Fourier-analytic arguments at the end of this section.

One important technical problem that needs addressing is that the function Λ is not bounded, which means that much of the analysis in previous sections, strictly speaking, does not apply. This is essentially equivalent to the fact that the primes have asymptotic density zero. However, one can resolve this problem by bounding Λ not by a bounded multiple of the constant function 1, which is not possible, but instead by a bounded multiple of another function ν which resembles Λ but is much easier to work with[8]. This corresponds to viewing the primes not as a (sparse) subset of the integers, but rather as a subset of the set of *almost primes*, which is much more tractable than the primes to study, and with the property that the primes have positive *relative* density inside the primes. One byproduct of this approach is that, because it uses very little about the primes other than this positive relative density, it in fact implies a stronger result, namely that *all* subsets of the primes with positive relative density must necessarily contain arbitrarily long arithmetic progressions.

Informally, the idea is as follows. Let P be the set of prime numbers between $N/2$ and N. The sieve of Eratosthenes shows that P consists precisely of those integers in $\{N/2, \ldots, N\}$ which are coprime to all primes less than $\sqrt{N}$. Motivated by this, let us define the partially sifted set P_R to be those integers in $\{N/2, \ldots, N\}$ which are coprime to all primes less than R, where $1 \leqslant R \leqslant \sqrt{N}$ is a parameter. Thus as R increases to $\sqrt{N}$, P_R decreases until it becomes P. The first few sets P_R are easy to understand, for instance P_2 is simply the odd numbers from $N/2$ to N. In particular, any statistic involving P_R (e.g. counting how many arithmetic progressions of length k are contained in P_R) is quite easy to compute to high accuracy when R is small. However, the task becomes increasingly difficult when R gets large. The vast and well-developed topic of *sieve theory* - a key component of analytic number theory - is devoted to questions like this; while this theory is too complex to be surveyed here, let us oversimplify one of the basic results in that field, namely the *fundamental lemma of sieve theory* In our notation, this lemma roughly speaking asserts that that one can compute the statistics

[8] As before we are ignoring some details concerning how one embeds Λ inside $\mathbb{Z}/N\mathbb{Z}$; also, it turns out to be convenient to "factor out" the initial σ-algebra $\mathcal{B}_w$ by passing to a single atom, such as the residue class $1 \mod W$; we ignore these minor technical issues here.

of P_R as long as R is a sufficiently small power of N. For instance, one can accurately count the number of arithmetic progressions in P_R of length k if R is less than $N^{1/2k}$.

An informal probabilistic argument suggests that

$$|P_R| \sim \frac{N}{2} \prod_{p<R} (1 - \frac{1}{p}) \sim \frac{N}{\log R}$$

where we use $X \sim Y$ to denote equivalence up to constants (i.e. $C^{-1}Y \leqslant X \leqslant CY$). A famous theorem of Merten in fact gives the more precise asymptotic

$$|P_R| = (1 + o(1)) \frac{N}{2} \frac{e^{-\gamma}}{\log R}$$

as long as R is much less than $\sqrt{N}$ but goes to infinity as $N \to \infty$. Here $\gamma = 0.577\ldots$ is Euler's constant. Comparing this with the prime number theorem

$$|P| = (1 + o(1)) \frac{N}{2} \frac{1}{\log N}$$

we see that P will have a relative density $|P|/|P_R|$ bounded away from zero as long as we set R to equal a small power of N, say $R = N^\varepsilon$ for some fixed ε (this ε will eventually depend on k; in [23] it is $\varepsilon = \frac{1}{k2^{k+4}}$).

A natural choice for the weight function ν would then be $\log R 1_{P_R}$; this function would thus be normalized to essentially have mean 1, and Λ would be dominated by a bounded multiple of ν. For technical reasons, however, the function 1_{P_R} is a bit too "rough" to serve as a good weight function, and it is better to use a slightly "smoother" variant of this function, namely the truncated divisor sums studied by Goldston and Yildirim [13], [14], [15]. These are formed by replacing the von Mangoldt function

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

with the variant

$$\Lambda_R(n) := \sum_{d|n} \mu(d) (\log \frac{R}{d})_+$$

where $x_+ := \max(x, 0)$ is the positive part of x. One can easily verify that Λ_R is equal to $\log R$ on the set of P_R, and can thus be thought of as the function $\log R 1_{P_R}$ with an additional "tail". The advantage of working with Λ_R instead of P_R is that Λ_R is easily expressed as a linear combination of the functions $1_{d|n}$, i.e. the characteristic functions of the residue class $d\mathbb{Z}$. Moreover, the coefficients $\mu(d)(\log \frac{R}{d})_+$ for this linear combination are supported on the small values of d, which are easier to control; this is roughly analogous in harmonic analysis to a function having Fourier transform supported on the "low frequencies", which explains why such functions in number theory are sometimes referred to as being "smooth". In particular, the work of Goldston and Yildirim showed that (providing R was a sufficiently small power of N) it was possible to accurately estimate such expressions as

$$\mathbb{E}(\Lambda_R(n)\Lambda_R(n + r) \ldots \Lambda_R(n + (k - 1)r)|n, r \in \mathbb{Z}/N\mathbb{Z}).$$

We cannot directly use Λ_R to dominate Λ, as it turns out to oscillate in sign; however this is easily fixed by using instead the function $\nu(n) := \frac{1}{\log R}\Lambda_R^2(n)$. Actually, this is an oversimplification; in practice we need to localize n to an arithmetic progression of spacing W and length equal to cN for a small multiple of N. After these adjustments, Goldston and Yildirim essentially showed that ν was "pseudorandom" - that almost all the correlations of ν were very close to 1 (a formal definition of this rather technical statement is in [23]). Another way of saying this is that ν lies very close to 1 in certain "weak" norms (such as the Gowers uniformity norms). With this pseudorandomness property, it turns out that the weight ν behaves very similarly to 1, thus for instance the generalized von Neumann theorem, Lemma 3.4, can be extended to the case where f is bounded by the pseudorandom function ν rather than the constant function 1 (although one has to accept some additional $o(1)$ errors when doing so). See [23] for details; the ideas here were initially motivated by similar arguments in the setting of hypergraphs by Gowers [18].

We can now describe the proof of (5.4) for general k. For sake of concreteness we shall restrict ourselves to the case $k = 4$, although the argument extends without difficulty to higher k. We shall use the machinery developed in the energy increment proof of Szemerédi's theorem in the $k = 4$ case.

As discussed earlier, the objective is to locate a σ-algebra $\mathcal{B}$ such that

$$\|\Lambda - \mathbb{E}(\Lambda|\mathcal{B})\|_{U^{k-1}} \text{ is small} \tag{7.3}$$

(where we shall be a bit vague as to what "small" means), and such that $\mathbb{E}(\Lambda|\mathcal{B})$ is bounded. The choice $\mathcal{B} = \mathcal{B}_w$, where w as before is a slowly growing function of N, will obey the second property (this is basically Dirichlet's theorem), but it is unknown as to whether it obeys the first property. Nevertheless, we can proceed by a stopping time argument, somewhat similar to the Calderón-Zygmund stopping time arguments used in harmonic analysis, or the stopping time argument used in the proof of the Szemerédi regularity lemma. The key point is that if (7.3) fails for some algebra $\mathcal{B}$, then by setting g to be the dual function of $\Lambda - \mathbb{E}(\Lambda|\mathcal{B})$,

$$g := \mathcal{D}(\Lambda - \mathbb{E}(\Lambda|\mathcal{B})),$$

then g will have a non-trivial correlation with $\Lambda - \mathbb{E}(\Lambda|\mathcal{B})$:

$$|\langle g, \Lambda - \mathbb{E}(\Lambda|\mathcal{B})\rangle| \text{ is large.}$$

Viewing this geometrically in the Hilbert space $L^2(\mathbb{Z}_N)$, this means that Λ (now thought of as a vector) contains a non-trivial component which is orthogonal to the subspace $L^2(\mathcal{B})$ which the conditional expectation operator $\mathbb{E}(|\mathcal{B})$ projects to, and which is also somewhat parallel to g. Thus if one defines $\mathcal{B}'$ to be the algebra generated by $\mathcal{B}$ and (suitable level sets of) g, we expect $L^2(\mathcal{B}')$ to capture both $L^2(\mathcal{B})$ and g (or a vector very close to g). Putting this together, we expect Λ to be closer to the subspace $L^2(\mathcal{B}')$ than to the smaller subspace $L^2(\mathcal{B})$; indeed, some applications of Cauchy-Schwarz and Pythagoras's theorem can be used to give an energy increment estimate of the form

$$\|\mathbb{E}(\Lambda|\mathcal{B}')\|_{L^2}^2 \geqslant \|\mathbb{E}(\Lambda|\mathcal{B})\|_{L^2}^2 + c \tag{7.4}$$

for some $c > 0$ (which depends of course on the definitions of "small" and "large").

To summarize, whenever (7.3) fails, we can exploit this failure to enlarge the underlying σ-algebra $\mathcal{B}$ in such a way that it collects more of the "energy" of Λ. We can now replace $\mathcal{B}$ by $\mathcal{B}'$ and iterate this procedure until (7.3) is finally attained. At first glance it seems that this algorithm could continue for quite a long time, since Λ has a large L^2 norm. Fortunately, though, it turns out that $\mathbb{E}(\Lambda|\mathcal{B})$ remains uniformly bounded throughout this algorithm. This is because Λ is bounded by ν, and thus

$\mathbb{E}(\Lambda|\mathcal{B})$ is bounded by $\mathbb{E}(\nu|\mathcal{B})$. The latter function turns out to be bounded because ν is pseudorandom (and thus very uniform), whereas $\mathcal{B}$ was essentially generated by dual functions (and thus highly non-uniform). Indeed, it turns out that even if one runs this algorithm for a large number of iterations, the bounds on $\mathbb{E}(\nu|\mathcal{B})$ only worsen by at most $o(1)$. This crucial fact is one of the more delicate computations in [23], but it ultimately follows from the pseudorandomness information on ν and an application of the Gowers-Cauchy-Schwarz inequality (3.3). This boundedness of $\mathbb{E}(\Lambda|\mathcal{B})$ is required for two reasons: firstly, in order that Szemerédi's theorem (in its third formulation) can be applied to this function, and secondly it is used (in conjunction with (7.4)) to show that the algorithm to find $\mathcal{B}$ halts after only a bounded number of iterations.

We now briefly remark on the earlier $k = 3$ versions of the above argument, referring the reader to [20], [24] for further details. In that case, the notion of pseudorandomness of the dominating measure ν was replaced by that of *linear pseudorandomness* or *Fourier pseudorandomness*, which basically asserts that all the Fourier coefficients of $\nu - 1$ were small. By Tomas-Stein restriction type arguments, this implies a certain Fourier restriction theorem for ν, which can be used to develop weighted analogues of Proposition 2.3 adapted to ν. One then runs the same argument as before, but this time the σ-algebra $\mathcal{B}$ is more explicit: it is the algebra generated by the Bohr sets corresponding to those frequencies where the Fourier transform of Λ is large. (Of course, the Hardy-Littlewood method already provides information as to where this Fourier transform is large; however the advantage of this argument is that it still works if Λ is replaced by any other function supported on a dense subset of the primes, whereas the Hardy-Littlewood method relies on the arithmetic structure on Λ and does not extend in this manner). Again, the pseudorandomness of ν will ensure that $\mathbb{E}(\nu|\mathcal{B})$, and hence $\mathbb{E}(\Lambda|\mathcal{B})$, is bounded, and one can then apply (the third version of) Roth's theorem to deduce Theorem 1.4. (Some further variations of this theme are pursued in [24]).

References

[1] I. Assani, *Pointwise convergence of ergodic averages along cubes*, preprint.

[2] A. Balog, *Linear equations in primes,* Mathematika **39** (1992) 367–378.

[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] J. Bourgain *A Szemerédi-type theorem for sets of positive density in* $\mathbb{R}^k$*,* Israel J. Math **54** (1986), no. 3, 307–316.

[5] ———, *On triples in arithmetic progression*, GAFA **9** (1999), 968–984.

[6] F.R.K. Chang, R. Graham, *Quasi-random subsets of* $\mathbb{Z}_n$, J. Comb. Th. A. **61** (1992), 64–86.

[7] S. Chowla, *There exists an infinity of 3—combinations of primes in A. P.*, Proc. Lahore Philos. Soc. **6**, (1944). no. 2, 15–16.

[8] M. Christ, A. Carbery, J. Wright, *Multidimensional van der Corput and Sublevel set estimates*, J. Amer. Math. Soc. **12** (1999), 981-1015.

[9] P. Erdős, P. Turán, *On some sequences of integers*, J. London Math. Soc. **11** (1936), 261–264.

[10] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.

[11] H. Furstenberg, Y. Katznelson and D. Ornstein, *The ergodic-theoretical proof of Szemerédi's theorem,* Bull. Amer. Math. Soc. **7** (1982), 527–552.

[12] H. Furstenberg, B. Weiss, *A mean ergodic theorem for* $1/N \sum_{n=1}^{N} f(T^n x) g(T^{n^2} x)$, Convergence in ergodic theory and probability (Columbus OH 1993), 193–227, Ohio State Univ. Math. Res. Inst. Publ., 5. de Gruyter, Berlin, 1996.

[13] D. Goldston and C.Y. Yildirim *Higher correlations of divisor sums related to primes, I: Triple correlations,* Integers **3** (2003) A5, 66pp.

[14] ———, *Higher correlations of divisor sums related to primes, III: k-correlations,* preprint (available at AIM preprints)
[15] ———, *Small gaps between primes, I,* preprint.
[16] T. Gowers, *A new proof of Szemerédi's theorem for arithmetic progressions of length four*, GAFA **8** (1998), 529–551.
[17] ———, *A new proof of Szemerédi's theorem*, GAFA **11** (2001), 465-588.
[18] ———, *Hypergraph regularity and the multidimensional Szemerédi theorem,* preprint
[19] R. Graham, B. Rothschild, J.H. Spencer, Ramsey Theory, John Wiley and Sons, NY (1980).
[20] B.J. Green, *Roth's theorem in the primes,* preprint.
[21] ———, *A Szemerédi-type regularity lemma in abelian groups,* preprint.
[22] ———, *Finite field models in arithmetic combinatorics*, preprint.
[23] B.J. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions,* preprint.
[24] B.J. Green and T. Tao, *Restriction theory of Selberg's sieve, with applications,* preprint.
[25] B.J. Green and T. Tao, *An inverse theorem for the Gowers U^3 norm*, preprint.
[26] G.H. Hardy and J.E. Littlewood *Some problems of "partitio numerorum"; III: On the expression of a number as a sum of primes,* Acta Math. **44** (1923), 1–70
[27] D.R. Heath-Brown, *Three primes and an almost prime in arithmetic progression,* J. London Math. Soc. (2) **23** (1981), 396–414.
[28] ———, *Linear relations amongst sums of two squares,* Number theory and algebraic geometry — to Peter Swinnerton-Dyer on his 75th birthday, CUP (2003).
[29] ———, *Non-conventional ergodic averages and nilmanifolds,* to appear in Ann. Math.
[30] M. Lacey, C. Thiele, *L^p estimates on the bilinear Hilbert transform for $2 < p < \infty$.* Ann. Math. 146 (1997), pp. 693–724.
[31] A. Moran, P. Pritchard and A. Thyssen, *Twenty-two primes in arithmetic progression,* Math. Comp. **64** (1995), no. 211, 1337–1339.
[32] O. Ramaré, *On Snirel'man's constant,* Ann. Scu. Norm. Pisa **21** (1995), 645–706.
[33] O. Ramaré and I.Z. Ruzsa, *Additive properties of dense subsets of sifted sequences,* J. Th. Nombres de Bordeaux **13** (2001) 559–581.
[34] V. Rödl, J. Skokan, *Regularity lemma for k-uniform hypergraphs*, to appear, Random Structures and Algorithms.
[35] V. Rödl, J. Skokan, *Applications of the regularity lemma for uniform hypergraphs*, preprint.
[36] K.F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 245-252.
[37] J. Solymosi, *A note on a question of Erdös and Graham*, Combinatorics, Probability and Computing **13** (2004), 263–267.
[38] E. Szemerédi, *On sets of integers containing no four elements in arithmetic progression*, Acta Math. Acad. Sci. Hungar. **20** (1969), 89–104.
[39] ———, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 299–345.
[40] ———, *Regular partitions of graphs,* in "Proc. Colloque Inter. CNRS" (J.-C. Bermond, J.-C. Fournier, M. Las Vergnas, D. Sotteau, eds.) (1978), 399–401.
[41] T. Tao, *A quantitative ergodic theory proof of Szemerédi's theorem*, preprint.
[42] E.C. Titchmarsh, *The theory of the Riemann zeta function,* Oxford University Press, 2nd ed, 1986.
[43] J.G. van der Corput, *Über Summen von Primzahlen und Primzahlquadraten,* Math. Ann. **116** (1939), 1–50.
[44] B.L. Van der Waerden, *Beweis einer Baudetschen Vermutung*, Nieuw. Arch. Wisk. **15** (1927), 212–216.
[45] P. Varnavides, *On certain sets of positive density,* J. London Math. Soc. **34** (1959) 358–360.
[46] I.M. Vinogradov, *Representation of an Odd Number as a Sum of Three Primes*, Comptes rendus (Doklady) de l'Académie des Sciences de l'U.R.S.S. **15** (1937a), 169–172.
[47] T. Ziegler, *Universal characteristic factors and Furstenberg averages*, preprint.
[48] ———, *A non-conventional ergodic theorem for a nilsystem*, preprint.

Department of Mathematics, University of California at Los Angeles, Los Angeles CA 90095
E-mail address: tao@math.ucla.edu