Current Developments in Mathematics, 2000
Pages 149 – 196

Arithmetic Progressions in Sparse Sets.

W. T. Gowers

1. Introduction.

The following statement is a famous theorem of van der Waerden.

Theorem 1.1. *Let k and r be positive integers. Then there exists a positive integer N such that, no matter how the set $\{1, 2, \dots, N\}$ is partitioned into r subsets $X_1 \cup \cdots \cup X_r$, at least one X_i contains an arithmetic progression of length k.*

It is customary to refer to the partition of $\{1, 2, \dots, N\}$ as an *r-colouring* and to the sets $X_1, \dots, X_r$ as *colour classes* or even simply *colours*. An arithmetic progression contained in one colour class is then known as *monochromatic*.

Over the years there has been considerable interest in the dependence of N on k and r. The proof of van der Waerden uses a double induction argument and the resulting bound is as bad as the Ackermann function, even when r is fixed at 2. The bound was not improved for several decades, and there was some speculation that it might even reflect the true state of affairs, until Shelah discovered a beautiful argument [Sh], surprisingly similar to the original one, which reduced the bound to a primitive recursive function. This function was still large, however: if you define T recursively by $T(1) = 1$ and $T(n+1) = 2^{T(n)}$, and define W recursively by $W(1) = 1$ and $W(n+1) = T(W(n))$, then W gives you some idea of the order of growth of Shelah's improved upper bound.

In 1936, Erdős and Turán conjectured the following dramatic strengthening of van der Waerden's theorem [ET], which was finally proved by Szemerédi almost four decades later [Sz1,2].

Theorem 1.2. *For every $\delta > 0$ and every positive integer k there exists a positive integer N such that every set $A \subset \{1, 2, \dots, N\}$ of size at least δN contains an arithmetic progression of length k.*

This is a strengthening since one can deduce Theorem 1.1 by setting $\delta = 1/r$. In other words, Szemerédi's theorem implies that in van der Waerden's theorem one can find an arithmetic progression not just in *some* colour class, but in the *largest* one. For this reason, Szemerédi's theorem is known as the *density version* of van der Waerden's theorem.

Erdős and Turán had two main reasons for making their conjecture. The first was that there seems to be no way to adapt the proof of van der Waerden's theorem to prove the density version. Thus, the density version needs a fundamentally new idea, and one can hope that this new idea will lead to much better bounds for van der Waerden's theorem. The second is that if one could prove the result for $N = \exp(\delta^{-1})$ (when δ is sufficiently small) then it would follow from the prime number theorem that the primes contain arithmetic progressions of every length.

Unfortunately, Szemerédi's proof used van der Waerden's theorem. As a result, he gave no new information about van der Waerden's theorem itself (at least from the point of view of bounds) and certainly not about arithmetic progressions in the primes. A further breakthrough was made by Furstenberg a few years later, who discovered a statement in ergodic theory which was equivalent to Szemerédi's theorem and proved the equivalent statement using ergodic-theoretic methods [Fu] (see also [FKO] for a simpler proof along similar lines). Furstenberg's approach led to a whole industry of powerful generalizations which is still continuing. However, it was non-constructive, so gave no information at all about bounds.

Indeed, the question about the primes is still an open problem, so the hope of Erdős and Turán has not yet been fulfilled. The situation regarding van der Waerden's theorem is different, however. The main topic of these lectures is a new proof of Szemerédi's theorem which, for the first time, gives the best known bound for van der Waerden's theorem as well. Our precise result is as follows. We shall write $a \uparrow b$ for a^b, with the obvious bracketing convention: $a \uparrow b \uparrow c$ means $a \uparrow (b \uparrow c)$.

Theorem 1.3. *Let $0 < \delta \leq 1/2$, let k be a positive integer, let $N \geq 2 \uparrow 2 \uparrow \delta^{-1} \uparrow 2 \uparrow 2 \uparrow (k+9)$ and let A be a subset of the set $\{1, 2, \dots, N\}$ of size at least δN. Then A contains an arithmetic progression of length k.* □

Corollary 1.4. *Let k be a positive integer and let $N \geq 2 \uparrow 2 \uparrow 2 \uparrow 2 \uparrow 2 \uparrow (k+9)$. Then however the set $\{1, 2, \dots, N\}$ is coloured with two colours, there will be a monochromatic arithmetic progression of length k.* □

The bound given by Corollary 1.4 is large, but it is two levels below Shelah's bound in the Ackermann hierarchy, and could be regarded as the first 'reasonable' bound for van der Waerden's theorem.

An additional interesting feature of our proof is that it uses methods of additive number theory, such as estimating exponential sums. Since many additive theorems about the primes have been proved using similar methods, there is a possibility that some combination of all the ideas may prove that the primes contain arbitrarily long arithmetic progressions. That is, it may be possible to solve the primes problem using more about the primes than just the prime number theorem. However, this is still some way off, and even the following question is unsolved: do the primes contain infinitely many arithmetic progressions of length four?

It is not be possible even to sketch the entire proof in an hour and a half - as it is written at present [G] it takes 129 pages. My main aim in this article is to explain the ideas from [G] for progressions of length three and four. Even for progressions of length three the result is interesting, and was first proved by Roth in 1952 [R]. Progressions of length four turn out to be significantly harder. This is true for all

known proofs: the usual experience seems to be that if you have proved the result for progressions of length four, then you will eventually prove it for all lengths, though it may take a lot of effort. After discussing progressions of length four, I shall give some idea of why progressions of length five are harder again, and of how to deal with the extra difficulties that arise. Beyond five, there are no interesting further difficulties: once again, this was true for Szemerédi and Furstenberg as well.

The next few sections of this paper are lifted directly from [G] and modified, sometimes a little and sometimes a lot. In general, this paper is more discursive and replaces some of the precise arguments of [G] by sketches which may be easier to read. In this way I have tried to cater for the reader who wishes to skip details, while providing enough precision for the more careful reader to get a good idea of the most important arguments. The final, quite long section is a survey of open problems related to the results and techniques of this paper.

2. Uniform Sets and Roth's Theorem.

This section begins with some notation and one or two concepts that are fundamental to the rest of the paper, and then uses these concepts to give Roth's proof (slightly rewritten) of his theorem (which is Szemerédi's theorem for progressions of length three). I have given this proof in full detail, and then followed it by a less detailed version of the same argument for those who prefer it.

It is not hard to prove that a random subset of the set $\{1, 2, \dots, N\}$ of cardinality δN contains, with high probability, roughly the expected number of arithmetic progressions of length k, that is, δ^k times the number of such progressions in the whole of $\{1, 2, \dots, N\}$. A natural approach to Szemerédi's theorem is therefore to try to show that random sets contain the fewest progressions of length k, which would then imply the theorem. In view of many other examples in combinatorics where random sets are extremal, this is a plausible statement, but unfortunately it is false. Indeed, if random sets were the worst, then the value of δ needed to ensure an arithmetic progression of length three would be of order of magnitude $N^{-2/3}$, whereas in fact it is known to be at least $\exp(-c(\log N)^{1/2})$ for some absolute constant $c > 0$ [Be]. We shall present this lower bound in the last section of the paper. (The random argument suggested above is to choose δ so that the expected number of arithmetic progressions is less than one. Using a standard trick in probabilistic combinatorics, we can instead ask for the expected number to be at most $\delta N/2$ and then delete one point from each one. This slightly better argument lifts the density significantly, but still only to $cN^{-1/2}$.)

Despite Behrend's example, it is tempting to try to exploit the fact that random sets contain long arithmetic progressions. Such a proof could be organized as follows.

(1) Define an appropriate notion of pseudorandomness.
(2) Prove that every pseudorandom subset of $\{1, 2, \dots, N\}$ contains roughly the number of arithmetic progressions of length k that you would expect.
(3) Prove that if $A \subset \{1, 2, \dots, N\}$ has size δN and is *not* pseudorandom, then there exists an arithmetic progression $P \subset \{1, 2, \dots, N\}$ with length tending to

infinity with N, such that $|A \cap P| \geq (\delta + \epsilon)|P|$, for some $\epsilon > 0$ that depends on δ (and k) only.

If these three steps can be carried out, then a simple iteration proves Szemerédi's theorem. As we shall see, this is exactly the scheme of Roth's proof for progressions of length three.

First, we must introduce some notation. Throughout the paper we shall be considering subsets of $\mathbb{Z}_N$ rather than subsets of $\{1, 2, \ldots, N\}$. It will be convenient (although not essential) to take N to be a prime number. We shall write ω for the number $\exp(2\pi i/N)$. Given a function $f : \mathbb{Z}_N \to \mathbb{C}$ and $r \in \mathbb{Z}_N$ we set

$$\hat{f}(r) = \sum_{s \in \mathbb{Z}_N} f(s)\omega^{-rs} .$$

The function $\hat{f}$ is the discrete Fourier transform of f. (In most papers in analytic number theory, the above exponential sum is written $\sum_{s=1}^{N} e(-rs/N)$, or possibly $\sum_{s=1}^{N} e_N(-rs)$.) Let us write $f * g$ for the function

$$f * g(s) = \sum_{t \in \mathbb{Z}_N} f(t)\overline{g(t-s)} .$$

(This is not standard notation, but we shall have no use for the convolution $\sum f(t)g(s-t)$ in this paper, so it is very convenient.) From now on, all sums will be over $\mathbb{Z}_N$ unless it is specified otherwise. We shall use the following basic identities over and over again in the paper.

$$(f * g)\hat{}(r) = \hat{f}(r)\overline{\hat{g}(r)} \tag{1}$$

$$\sum_r \hat{f}(r)\overline{\hat{g}(r)} = N \sum_s f(s)\overline{g(s)} \tag{2}$$

$$\sum_r |\hat{f}(r)|^2 = N \sum_s |f(s)|^2 \tag{3}$$

$$f(s) = N^{-1} \sum_r \hat{f}(r)\omega^{rs} \tag{4}$$

Of these, the first tells us that convolutions transform to pointwise products, the second and third are Parseval's identities and the last is the inversion formula. To check them directly, note that

$$\begin{aligned}(f * g)(r) &= \textstyle\sum_s (f * g)(s)\omega^{-rs} \\ &= \textstyle\sum_{s,t} f(t)\overline{g(t-s)}\omega^{-rt}\omega^{r(t-s)} \\ &= \textstyle\sum_{t,u} f(t)\omega^{-rt}\overline{g(u)\omega^{-ru}} \\ &= \hat{f}(r)\overline{\hat{g}(r)}\end{aligned}$$

which proves (1). We may deduce (2), since

$$\sum_r \hat{f}(r)\overline{\hat{g}(r)} = \sum_r \sum_s f * g(s)\omega^{-rs} = Nf * g(0) = N \sum_s f(s)\overline{g(s)} ,$$

where for the second equality we used the fact that $\sum_s \omega^{-rs}$ is N if $r = 0$ and zero otherwise. Identity (3) is a special case of (2). Noting that the function $r \mapsto \omega^{-rs}$ is the Fourier transform of the characteristic function of the singleton $\{s\}$, we can deduce (4) from (2) as well (though it is perhaps more natural just to expand the right hand side and give a direct proof).

There is one further identity, sufficiently important to be worth stating as a lemma.

Lemma 2.1. *Let f and g be functions from $\mathbb{Z}_N$ to $\mathbb{C}$. Then*

$$\sum_r |\hat{f}(r)|^2 |\overline{\hat{g}(r)}|^2 = N \sum_t \Big| \sum_s f(s)\overline{g(s-t)} \Big|^2 . \tag{5}$$

Proof. By identities (1) and (2),

$$\begin{aligned} \sum_r |\hat{f}(r)|^2 |\overline{\hat{g}(r)}|^2 &= \textstyle\sum_r |(f*g)\hat{}(r)|^2 \\ &= N \textstyle\sum_t |f*g(t)|^2 \\ &= N \textstyle\sum_{t,s,u} f(s)\overline{g(s-t)}f(u)g(u-t) \\ &= N \textstyle\sum_t \Big| \sum_s f(s)\overline{g(s-t)} \Big|^2 \end{aligned}$$

as required. □

Setting $f = g$ and expanding the right hand side of (5), one obtains another identity which shows that sums of fourth powers of Fourier coefficients have an interesting interpretation.

$$\sum_r |\hat{f}(r)|^4 = N \sum_{a-b=c-d} f(a)\overline{f(b)f(c)}f(d) . \tag{6}$$

It is of course easy to check this identity directly.

Nearly all the functions in this paper will take values with modulus at most one. In such a case, one can think of Lemma 2.1 as saying that if f has a large inner product with a large number of rotations of g, then f and g must have large Fourier coefficients in common, where large means of size proportional to N. We shall be particularly interested in the Fourier coefficients of characteristic functions of sets $A \subset \mathbb{Z}_N$ of cardinality δN, which we shall denote by the same letter as the set itself. Notice that identity (6), when applied to (the characteristic function of) a set A, tells us that the sum $\sum_r |\hat{A}(r)|^4$ is N times the number of quadruples $(a,b,c,d) \in A^4$ such that $a - b = c - d$.

For technical reasons it is also useful to consider functions of mean zero. Given a set A of cardinality δN, let us define the *balanced* function of A to be $f_A : \mathbb{Z}^N \to [-1,1]$ where

$$f_A(s) = \begin{cases} 1-\delta & \text{s} \in \text{A} \\ -\delta & \text{s} \notin \text{A} \end{cases} .$$

This is the characteristic function of A minus the constant function $\delta\mathbf{1}$. Note that $\sum_{s\in\mathbb{Z}_N} f_A(s) = \hat{f}_A(0) = 0$ and that $\hat{f}_A(r) = \hat{A}(r)$ for $r \neq 0$.

We are now in a position to define a useful notion of pseudorandomness. The next lemma (which is not new) gives several equivalent definitions involving constants c_i. When we say that one property involving c_i implies another involving c_j, we mean that if the first holds, then so does the second for a constant c_j that tends to zero as c_i tends to zero. (Thus, if one moves from one property to another and

then back again, one does not necessarily recover the original constant.) From the point of view of the eventual bounds obtained, it is important that the dependence is no worse than a fixed power. This is always true below.

In this paper we shall use the letter D to denote the closed unit disc in $\mathbb{C}$ (unless it obviously means something else).

Lemma 2.2. *Let f be a function from $\mathbb{Z}_N$ to D. The following are equivalent.*
(i) $\sum_k \left|\sum_s f(s)\overline{f(s-k)}\right|^2 \leq c_1 N^3$.
(ii) $\sum_{a-b=c-d} f(a)\overline{f(b)f(c)}f(d) \leq c_1 N^3$.
(iii) $\sum_r |\hat{f}(r)|^4 \leq c_1 N^4$.
(iv) $\max_r |\hat{f}(r)| \leq c_2 N$.
(v) $\sum_k \left|\sum_s f(s)\overline{g(s-k)}\right|^2 \leq c_3 N^2 \|g\|_2^2$ *for every function* $g : \mathbb{Z}_N \to \mathbb{C}$.

Proof. The equivalence of (i) and (ii) comes from expanding the left hand side of (i), and the equivalence of (i) and (iii) follows from identity (6) above. It is obvious that (iii) implies (iv) if $c_2 \geq c_1^{1/4}$. Since

$$\sum_r |\hat{f}(r)|^4 \leq \max_r |\hat{f}(r)|^2 \sum_r |\hat{f}(r)|^2 \leq N^2 \max_r |\hat{f}(r)|^2 ,$$

we find that (iv) implies (iii) if $c_1 \geq c_2^2$. It is obvious that (v) implies (i) if $c_1 \geq c_3$. By Lemma 2.1, the left hand side of (v) is

$$N^{-1} \sum_r |\hat{f}(r)|^2 |\hat{g}(r)|^2 \leq N^{-1} \Big(\sum_r |\hat{f}(r)|^4\Big)^{1/2} \Big(\sum_r |\hat{g}(r)|^4\Big)^{1/2}$$

by the Cauchy-Schwarz inequality. Using the additional inequality

$$\Big(\sum_r |\hat{g}(r)|^4\Big)^{1/2} \leq \sum_r |\hat{g}(r)|^2 ,$$

we see that (iii) implies (v) if $c_3 \geq c_1^{1/2}$. □

A function $f : \mathbb{Z}_N \to D$ satisfying condition (i) above, with $c_1 = \alpha$, will be called *α-uniform*. If f is the balanced function f_A of some set $A \subset \mathbb{Z}_N$, then we shall also say that A is α-uniform. If $A \subset \mathbb{Z}_N$ is an α-uniform set of cardinality δN, and f is its balanced function, then

$$\sum_r |\hat{A}(r)|^4 = |A|^4 + \sum_r |\hat{f}(r)|^4 \leq |A|^4 + \alpha N^4 .$$

We noted earlier that $\sum_r |\hat{A}(r)|^4$ is N times the number of quadruples $(a,b,c,d) \in A^4$ such that $a-b=c-d$. If A were a random set of size δN, then we would expect about $\delta^4 N^3 = N^{-1}|A|^4$ such quadruples (which from the above is clearly a lower bound). Therefore, the number α is measuring how close A is to being random in this particular sense. Notice that quadruples (a,b,c,d) with $a-b=c-d$ are the same as quadruples of the form $(x, x+s, x+t, x+s+t)$.

We remark that our definition of an α-uniform set coincides with the definition of *quasirandom* subsets of $\mathbb{Z}_N$, due to Chung and Graham [CG]. They prove that several formulations of the definition (including those of this paper) are equivalent. They do not mention the connection with Roth's theorem, which we shall now explain. We need a very standard lemma, which we prove in slightly greater

generality than is immediately necessary, so that it can be used again later. Let us define the *diameter* of a subset $X \subset \mathbb{Z}_N$ to be the smallest integer s such that $X \subset \{n, n+1, \ldots, n+s\}$ for some $n \in \mathbb{Z}_N$.

Lemma 2.3. *Let r, s and N be positive integers with $r, s \leq N$ and $rs \geq N$, and let $\phi : \{0, 1, \ldots, r-1\} \to \mathbb{Z}_N$ be linear (i.e., of the form $\phi(x) = ax + b$). Then the set $\{0, 1, \ldots, r-1\}$ can be partitioned into arithmetic progressions $P_1, \ldots, P_M$ such that for each j the diameter of $\phi(P_j)$ is at most s and the length of P_j lies between $(rs/4N)^{1/2}$ and $(rs/N)^{1/2}$.*

Proof. Let $t = \lceil (rN/4s)^{1/2} \rceil$. Of the numbers $\phi(0), \phi(1), \ldots, \phi(t)$, at least two must be within N/t. Therefore, by the linearity of ϕ, we can find a non-zero $u \leq t$ such that $|\phi(u) - \phi(0)| \leq N/t$. Split $\{0, 1, \ldots, r-1\}$ into congruence classes mod u. Each congruence class is an arithmetic progression of cardinality either $\lfloor r/u \rfloor$ or $\lceil r/u \rceil$. If P is any set of at most st/N consecutive elements of a congruence class, then $\mathrm{diam}\phi(P) \leq s$. It is easy to check first that $st/N \leq r/3t \leq (1/2)\lfloor r/u \rfloor$, next that this implies that the congruence classes can be partitioned into sets P_j of consecutive elements with every P_j of cardinality between $\lceil st/2N \rceil$ and $\lfloor st/N \rfloor$, and finally that this proves the lemma. □

Corollary 2.4. *Let f be a function from the set $\{0, 1, \ldots, r-1\}$ to the closed unit disc in $\mathbb{C}$, let $\phi : \mathbb{Z}_N \to \mathbb{Z}_N$ be linear and let $\alpha > 0$. If*

$$\Big| \sum_{x=0}^{r-1} f(x) \omega^{-\phi(s)} \Big| \geq \alpha r$$

then there is a partition of $\{0, 1, \ldots, r-1\}$ into $m \leq (8\pi r/\alpha)^{1/2}$ arithmetic progressions $P_1, \ldots, P_m$ such that

$$\sum_{j=1}^{m} \Big| \sum_{x \in P_j} f(x) \Big| \geq (\alpha/2) r$$

and such that the lengths of the P_j all lie between $(\alpha r/\pi)^{1/2}/4$ and $(\alpha r/\pi)^{1/2}/2$.

Proof. Let $s \leq \alpha N/4\pi$ and let $m = (16\pi r/\alpha)^{1/2}$. By Lemma 2.3 we can find a partition of $\{0, 1, \ldots, r-1\}$ into arithmetic progressions $P_1, \ldots, P_m$ such that the diameter of $\phi(P_j)$ is at most s for every j and the length of each P_j lies between r/m and $2r/m$. By the triangle inequality,

$$\sum_{j=1}^{m} \Big| \sum_{x \in P_j} f(x) \omega^{-\phi(x)} \Big| \geq \alpha r \ .$$

Let $x_j \in P_j$. The estimate on the diameter of $\phi(P_j)$ implies that $|\omega^{-\phi(x)} - \omega^{-\phi(x_j)}|$ is at most $\alpha/2$ for every $x \in P_j$. Therefore

$$\begin{aligned} \sum_{j=1}^{m} \Big| \sum_{x \in P_j} f(x) \Big| &= \textstyle\sum_{j=1}^{m} \Big| \sum_{x \in P_j} f(x) \omega^{-\phi(x_j)} \Big| \\ &\geq \textstyle\sum_{j=1}^{m} \Big| \sum_{x \in P_j} f(x) \omega^{-\phi(x)} \Big| - \sum_{j=1}^{m} (\alpha/2) |P_j| \\ &\geq \alpha r/2 \end{aligned}$$

as claimed. □

Corollary 2.5. *Let $A \subset \mathbb{Z}_N$ and suppose that $|\hat{A}(r)| \geq \alpha N$ for some $r \neq 0$. Then there exists an arithmetic progression $P \subset \{0,1,\dots,N-1\}$ of length at least $(\alpha^3 N/128\pi)^{1/2}$ such that $|A \cap P| \geq (\delta + \alpha/8)|P|$.*

Proof. Define $\phi(x) = rx$ and let f be the balanced function of A (regarded as a function on $\{0,1,\dots,N-1\}$). By Corollary 2.4 we can partition the set $\{0,1,\dots,N-1\}$ into $m \leq (16\pi N/\alpha)^{1/2}$ arithmetic progressions $P_1,\dots,P_m$ of lengths between N/m and $2N/m$ such that

$$\sum_{j=1}^{m} \Big|\sum_{x\in P_j} f(x)\Big| \geq \alpha N/2 .$$

Since $\sum_{x\in P_j} f(x)$ is real for all j, and since $\sum_{j=1}^{m}\sum_{x\in P_j} f(x) = 0$, if we define J to be the set of j with $\sum_{x\in P_j} f(x) \geq 0$, we have

$$\sum_{j\in J}\sum_{x\in P_j} f(x) \geq \alpha N/4 .$$

Therefore, we can find j such that $\sum_{x\in P_j} f(x) \geq \alpha N/4m$. But $|P_j| \leq 2N/m$, so $\sum_{x\in P_j} f(x) \geq \alpha|P_j|/8$, which implies that $|A\cap P_j| \geq (\delta+\alpha/8)|P_j|$. □

We can now give Roth's proof of his theorem on arithmetic progressions of length three.

Theorem 2.6. *Let $\delta > 0$, let $N \geq \exp\exp(C\delta^{-1})$ (where C is an absolute constant) and let $A \subset \{1,2,\dots,N\}$ be a set of size at least δN. Then A contains an arithmetic progression of length three.*

Proof. Since we are passing to smaller progressions and iterating, we cannot simply assume that N is prime, so we shall begin by dealing with this small technicality. Let N_0 be a positive integer and let A_0 be a subset of $\{1,2,\dots,N_0\}$ of size at least $\delta_0 N_0$.

By Bertrand's postulate (which is elementary - it would be a pity to use the full strength of the prime number theorem in a proof of Roth's theorem) there is a prime p between $N/3$ and $2N/3$. Write q for $N-p$. If $|A_0\cap\{1,2,\dots,p\}| \leq \delta_0(1-\delta_0/160)p$, then we know that

$$|A_0\cap\{p+1,\dots,N\}| \geq \delta_0(N-(1-\delta_0/160)p) = \delta_0(q+\delta_0 p/160) \geq \delta_0(1+\delta_0/320)q .$$

Let us call this situation case 0.

If case 0 does not hold, then let N be the prime p obtained above, let $A = A_0\cap\{1,\dots,N\}$ and let $\delta = \delta_0(1-\delta_0/160)$. Let $B = A\cap[N/3,2N/3)$. If $|B| \leq \delta N/5$, then either $A\cap[0,N/3)$ or $A\cap[2N/3,N)$ has cardinality at least $2\delta N/5 = (6\delta/5)(N/3)$. This situation we shall call case 1.

Next, let $\alpha = \delta^2/10$ and suppose that $|\hat{A}(r)| > \alpha N$ for some non-zero r. In this case, by Corollary 2.5 there is an arithmetic progression P of cardinality at least $(\alpha^3 N/128\pi)^{1/2}$ such that $|A \cap P| \geq (\delta + \delta^2/80)|P|$. This situation will be case 2.

If case 2 does not hold, then $|\hat{A}(r)| \leq \alpha N$ for every non-zero r, which says that A satisfies condition (iv) of Lemma 2.2. The number of triples $(x, y, z) \in A \times B^2$ such that $x + z = 2y$ is then

$$N^{-1} \sum_{x \in A} \sum_{y \in B} \sum_{z \in B} \sum_r \omega^{r(2y-x-z)} = N^{-1} \sum_r \hat{A}(r)\hat{B}(-2r)\hat{C}(r)$$
$$\geq N^{-1}|A||B|^2 - N^{-1} \max_{r \neq 0} |\hat{A}(r)| \Big(\sum_{r \neq 0} |\hat{B}(-2r)|^2 \Big)^{1/2} \Big(\sum_{r \neq 0} |\hat{B}(r)|^2 \Big)^{1/2}$$
$$\geq \delta |B|^2 - \alpha |B| N \ .$$

If in addition case 1 does not hold, then this quantity is minimized when $|B| = \delta N/5$, and the minimum value is $\delta^3 N^2/50$, implying the existence of at least this number of triples $(x, y, z) \in A \times B^2$ in arithmetic progression mod N. Since B lives in the middle third, these are genuine progressions in $\{1, 2, \ldots, N\}$ and since there are only N degenerate progressions (i.e., with difference zero) we can conclude that A contains an arithmetic progression of length three as long as $N \geq 50\delta^{-3}$. This we shall call case 3.

To summarize, if case 3 holds and $N \geq 50\delta^{-3}$, then A contains an arithmetic progression of length three. In case 2, we can find a subprogression P of $\{1, \ldots, N\}$ of cardinality at least $(\alpha^3 N/128\pi)^{1/2}$ such that $|A \cap P| \geq \delta(1 + \delta/80)|P|$. Since $\{1, \ldots, N\}$ is a subprogression of $\{1, \ldots, N_0\}$, $A = A_0 \cap \{1, \ldots, N\}$ and one can easily check that $\delta(1 + \delta/80) \geq \delta_0(1 + \delta_0/320)$, we may conclude that in case 2 there is a subprogression P of $\{1, \ldots, N_0\}$ of cardinality at least $(\alpha^3 N_0/384\pi)^{1/2}$ such that $|A_0 \cap P| \geq \delta_0(1 + \delta_0/320)|P|$. As for cases 0 and 1, it is easy to see that the same conclusion also holds, and indeed a much stronger one as P has a length which is linear in N_0.

This gives us the basis for an iteration argument. If A_0 does not contain an arithmetic progression of length three, then we drop down to a progression P where the density of A is larger and repeat. If the density at step m of the iteration is δ_m, then at each subsequent iteration the density increases by at least $\delta_m^2/320$. It follows that the density reaches $2\delta_m$ after at most $320\delta_m^{-1}$ further steps. It follows that the total number of steps cannot be more than $320(\delta^{-1} + (2\delta)^{-1} + (4\delta)^{-1} + \ldots) = 640\delta^{-1}$. At each step, the size of the progression in which A lives is around the square root of what it was at the previous step. The result now follows from a simple calculation (left to the reader). □

Before we move to the next section, here is a sketch of the above argument, for the benefit of those who do not have the time or inclination to follow the details.

A sketched version of the above argument.

Let A be a subset of $\{1,2,\dots,N\}$ of size at least δN. It is convenient to identify $\{1,2,\dots,N\}$ with $\mathbb{Z}_N$ because the group structure of $\mathbb{Z}_N$ allows us to define discrete Fourier coefficients (see the precise definition at the beginning of this section). This is very useful, because the number of triples (x,y,z) in A^3 such that $x+z=2y \bmod N$ turns out to be $N^{-1}\sum_r \hat{A}(r)\hat{A}(-2r)\hat{A}(r)$. To see this, expand out the Fourier coefficients to obtain

$$N^{-1}\sum_r \sum_{x,y,z} A(x)\omega^{-rx}A(y)\omega^{2ry}A(z)\omega^{-rz} = N^{-1}\sum_{x,y,z} A(x)A(y)A(z)\sum_r \omega^{-r(x-2y+z)}.$$

The sum over r on the right hand side gives N if $x+z=2y \bmod N$ and otherwise vanishes, which proves the assertion.

Unfortunately, this raises a technical problem, since not every triple (x,y,z) with $x+z=2y \bmod N$ corresponds to an arithmetic progression in $\{1,2,\dots,N\}$. To get round this, it is convenient to look at three different sets A, B and C and count the number of triples $(x,y,z)\in A\times B\times C$ such that $x+z=2y \bmod N$. If B and C are subsets of the interval $[N/3,2N/3]$, then any such triple must correspond to a genuine arithmetic progression (apart from the degenerate triples (x,x,x) but there are few enough of these that they will not cause us any trouble).

Exactly as above, the number of triples in question can be interpreted on the Fourier side as $N^{-1}\sum_r \hat{A}(r)\hat{B}(-2r)\hat{C}(r)$. We then divide this sum into two parts, the contribution from $r=0$ and the rest. Now $N^{-1}\hat{A}(0)\hat{B}(0)\hat{C}(0) = |A||B||C|$. Writing $|A|=\alpha N$, $|B|=\beta N$ and $|C|=\gamma N$, we see that this is $\alpha\beta\gamma N^2$, which can be thought of as the expected number of progressions (mod N) if the elements of A are chosen randomly and independently with probability α, and similarly for B and C.

Standard probabilistic arguments show that with very high probability the number of mod-N progressions in $A\times B\times C$ will be very close to $\alpha\beta\gamma N^2$, so we can think of the rest of the sum as the *deviation from randomness* of the triple (A,B,C). In particular, if we can show that $N^{-1}\sum_{r\neq 0}\hat{A}(r)\hat{B}(-2r)\hat{C}(r)$ is at most $\alpha\beta\gamma N/2$, then we will have shown that $A\times B\times C$ contains many mod-N arithmetic progressions.

Of course, it is easy to think of examples of triples (A,B,C) such that $A\times B\times C$ contains no mod-N arithmetic progressions, so we cannot prove this unconditionally. However, we can prove it under the assumption that $\max_{r\neq 0}\hat{A}(r)\leq cN$ for a sufficiently small constant c (depending on α, β and γ). To do this, one simply applies the Cauchy-Schwarz inequality and Parseval's identity to say

$$\begin{aligned} N^{-1}\sum_{r\neq 0}\hat{A}(r)\hat{B}(-2r)\hat{C}(r) &\leq c\sum_r \hat{B}(-2r)\hat{C}(r) \\ &\leq c\Big(\sum_r |\hat{B}(-2r)|^2\Big)^{1/2}\Big(\sum_r |\hat{C}(r)|^2\Big)^{1/2} \\ &= c(N|B|)^{1/2}(N|C|)^{1/2} = c(\beta\gamma)^{1/2}N^2. \end{aligned}$$

Hence, for the argument to work, one needs $c(\beta\gamma)^{1/2}\leq\alpha\beta\gamma$, or $c\leq\alpha(\beta\gamma)^{1/2}$.

What happens if there exists r such that $|\hat{A}(r)|>cN$ for this c? Then we rely on a different argument. The statement that A has a non-trivial large Fourier coefficient at r is the statement that the sum $\sum_{x\in A}\omega^{-rx}$ is large, which means

that the values rx for $x \in A$ are not evenly distributed in $\mathbb{Z}_N$. This implies that there is a long interval $I = \{k, k+1, \dots, l\}$ such that $|rA \cap I|$ is significantly larger than $\alpha|I|$, which implies that $|A \cap r^{-1}I|$ is significantly larger than $|r^{-1}I| = |I|$. (The length of I is proportional to N.)

The set $r^{-1}I$ is an arithmetic progression mod N with common difference r^{-1}. A standard application of the pigeonhole principle (see Lemma 2.3 for the method of proof) allows us to partition I into about $\sqrt{N}$ subsets corresponding to genuine subprogressions of $\{1, 2, \dots, N\}$ and then, by an easy averaging argument, to find one of these, P say, such that $|P|$ is about $\sqrt{N}$ and $|A \cap P|$ is significantly larger than $\alpha|P|$. In fact, the proof gives $|A \cap P| \geq \alpha(1 + c_1\alpha)|P|$ for an absolute constant $c_1 > 0$.

This gives us the basis for an iteration argument along the lines sketched at the beginning of this section. Either A is random-like, meaning that it has no unexpectedly large Fourier coefficients, in which case it contains plenty of arithmetic progressions of length three, or it isn't, in which case we can use a large Fourier coefficient to find a longish subprogression where A has increased density. We can then restrict our attention to this subprogression and repeat the argument. Eventually, the density reaches 1 and the iteration cannot continue. It can be checked that this argument implies that any subset of $\{1, 2, \dots, N\}$ of size at least $CN/\log\log N$ contains an arithmetic progression of length three.

I have ignored two technicalities: N may not be prime (this becomes important when we iterate) and B and C, which we set to be $A \cap [N/3, 2N/3]$, may be very small. However, these are easily dealt with, as can be seen in the detailed proof of Theorem 2.6.

3. Higher-Degree Uniformity.

As I have said, the scheme of proof of Roth's theorem can be summarized as follows: if A is random-like, then it contains many progressions of length three. If, on the other hand, it is not random-like, then there is a reasonably long subprogression inside which A has increased density, making possible an iterative argument. A set A counts as random-like for this argument if it is α-uniform for some small α (the definition appears just after Lemma 2.2), which means, amongst other things, that $\hat{A}(r)$ is small for every non-zero r.

As will be shown in the next section, α-uniformity is not an appropriate definition of pseudorandomness when we come to look at progressions of length greater than three, because it turns out not to imply all that much about the number of such progressions. The main purpose of this section is therefore to introduce a stronger definition which can be used instead. I shall do so informally to begin with, and state the main result that is needed. The rest of the section contains a complete proof of the result.

Recall from the remarks following Lemma 2.2 that a set A of cardinality δN is α-uniform for a small α if and only if the number of quadruples $(x, x+a, x+b, x+a+b) \in A^4$ is not much larger than the lower bound of $\delta^4 N^4$, which holds for all such sets A and is more or less attained by random ones. The quadruples $(x, x+a, x+b, x+a+b)$ can be thought of as two-dimensional Hilbert cubes, and to

produce a definition of pseudorandomness suitable for progressions of length four we simply increase this dimension to three. Accordingly, we shall say that a set A of size δN is pseudorandom (the term we shall actually use is *quadratically uniform*) if the number of octuples

$$(x, x+a, x+b, x+c, x+a+b, x+a+c, x+b+c, x+a+b+c) \in A^8$$

is not much more than $\delta^8 N^4$, which, once again, can be shown quite easily to be a lower bound, almost attained by random sets.

The result needed later will be that if A is quadratically uniform in this sense (of course, to make this precise it is necessary to make the definition quantitative, but this is easily done) then A contains roughly the expected number of progressions of length four that will be contained in a random set of the same cardinality. In other words, when we talk about progressions of length four, quadratic uniformity is an appropriate definition of pseudorandomness. Moreover, the obvious generalization of this idea is the correct one: if we define a set A to be uniform of degree k when it contains roughly $\delta^{2^{k+1}} N^{k+2}$ cubes of dimension $k+1$, then, as one might hope, a set that is sufficiently uniform of degree k will contain roughly the expected number of progressions of length $k+2$.

Of course, these definitions will not help us unless we can go on to say something useful about sets that *fail* to be quadratically uniform or uniform of higher degree. Since quadratic uniformity is a stronger condition than uniformity, one would expect this to be harder than the corresponding step in Roth's theorem - proving that a set with a large Fourier coefficient can be restricted to a subprogression where it has increased density. And indeed it is, much harder. This is where most of the work is needed, and where the main interest in the proof lies.

It is worth mentioning an equivalent definition of quadratic uniformity, which relates it to Fourier coefficients. A set A is α-uniform if $\hat{A}(r)$ is at most αN whenever r is non-zero. It is quadratically uniform if, for almost every $k \in \mathbb{Z}_N$, the set $A \cap (A+k) = \{x : x \in A, x-k \in A\}$ is uniform. More precisely, one could say that A is α-quadratically uniform if the number of k for which $A \cap (A+k)$ fails to be α-uniform is at most αN. (Similarly, a set is cubically uniform if almost all the sets $A \cap (A+k)$ are quadratically uniform, or equivalently if almost all the sets $A \cap (A+k) \cap (A+l) \cap (A+k+l)$ are uniform.)

The rest of this section makes the above ideas precise and contains a proof that appropriately uniform sets contain appropriately many arithmetic progressions. For technical reasons we shall actually work with *functions* rather than sets, so the definitions below do not quite coincide with the one I have given. However, the reader who wishes to skip technical details ought to be able to jump to the next section and still understand the rest of the paper reasonably well.

In order to simplify the presentation for the case of progressions of length four, we shall now prove two lemmas, even though the second implies the first. Given a function $f : \mathbb{Z}_N \to \mathbb{Z}_N$, we shall define, for any k, the *difference function* $\Delta(f;k)$ by $\Delta(f;k)(s) = f(s)\overline{f(s-k)}$. The reason for the terminology is that if, as will often be the case, $f(s) = \omega^{\phi(s)}$ for some function $\phi : \mathbb{Z}_N \to \mathbb{Z}_N$, then $\Delta(f;k)(s) = \omega^{\phi(k)-\phi(s-k)}$.

Now let us define iterated difference functions in two different ways as follows. The first is inductive, setting $\Delta(f;a_1,\dots,a_d)(s)$ to be $\Delta(\Delta(f;a_1,\dots,a_{d-1});a_d)(s)$. The second makes explicit the result of the inductive process. Let C stand for the map from $\mathbb{C}^N$ to $\mathbb{C}^N$ which takes a function to its pointwise complex conjugate.

Given a function $f : \mathbb{Z}_N \to \mathbb{C}$, we define

$$\Delta(f; a_1, \dots, a_d)(s) = \prod_{\epsilon_1,\dots,\epsilon_d} (C^{\epsilon_1+\dots+\epsilon_d} f)\Big(s - \sum_{i=1}^{d} a_i\epsilon_i\Big)$$

where the product is over all sequences $\epsilon_1, \dots, \epsilon_d$ with $\epsilon_i \in \{0, 1\}$. When $d = 3$, for example, this definition becomes

$$\begin{aligned}\Delta(f; a, b, c)(s) &= f(s)\overline{f(s-a)f(s-b)f(s-c)}\times \\ &\times f(s-a-b)f(s-a-c)f(s-b-c)\overline{f(s-a-b-c)}\,.\end{aligned}$$

We now define a function f from $\mathbb{Z}_N$ to the closed unit disc $D \subset \mathbb{C}$ to be *α-uniform of degree d* if

$$\sum_{a_1,\dots,a_d} \Big|\sum_s \Delta(f; a_1, \dots, a_d)(s)\Big|^2 \le \alpha N^{d+2}\,.$$

When d equals two or three, we say that f is quadratically or cubically α-uniform respectively. As with the definition of α-uniformity (which is the same as α-uniformity of degree one) this definition has several useful reformulations.

Lemma 3.1. *Let f be a function from $\mathbb{Z}_N$ to D. The following are equivalent.*
(i) f is c_1-uniform of degree d.
(ii) $\sum_s \sum_{a_1,\dots,a_{d+1}} \Delta(f; a_1, \dots, a_{d+1})(s) \le c_1 N^{d+2}$.
(iii) There is a function $\alpha : \mathbb{Z}_N^{d-1} \to [0, 1]$ such that $\sum_{a_1,\dots,a_{d-1}} \alpha(a_1, \dots, a_{d-1}) \le c_1 N^{d-1}$ and $\Delta(f; a_1, \dots, a_{d-1})$ is $\alpha(a_1, \dots, a_{d-1})$-uniform for every $(a_1, \dots, a_{d-1})$.
(iv) There is a function $\alpha : \mathbb{Z}_N \to [0, 1]$ such that $\sum_r \alpha(r) = c_1 N$ and $\Delta(f; r)$ is $\alpha(r)$-uniform of degree $d-1$ for every r.
(v) $\sum_{a_1,\dots,a_{d-1}} \sum_r \big|\Delta(f; a_1, \dots, a_{d-1})\hat{}(r)\big|^4 \le c_1 N^{d+3}$.
(vi) For all but $c_2 N^{d-1}$ choices of $(a_1, \dots, a_{d-1})$ the function $\Delta(f; a_1, \dots, a_{d-1})$ is c_2-uniform.
(vii) There are at most $c_3 N^{d-1}$ values of $(a_1, \dots, a_{d-1})$ for which there exists some $r \in \mathbb{Z}_N$ with $\big|\Delta(f; a_1, \dots, a_{d-1})\hat{}(r)\big| \ge c_3 N$.

Proof. The equivalence of (i) and (ii) is easy, as the left hand sides of the relevant expressions are equal. It is also obvious that (ii) and (iii) are equivalent. A very simple inductive argument shows that (ii) is equivalent to (iv). The equivalence of (i) and (v) follows, as in the proof of the equivalence of (i) and (iii) in Lemma 2.1, by expanding the left hand side of (v). Alternatively, it can be deduced from Lemma 2.1 by applying that equivalence to each function $\Delta(f; a_1, \dots, a_{d-1})$ and adding.

Averaging arguments show that (iii) implies (vi) as long as $c_1 \le c_2^2$, and that (vi) implies (iii) as long as $c_1 \ge 2c_2$. Finally, the equivalence of (i) and (ii) in Lemma 2.1 shows that in this lemma (vi) implies (vii) if $c_3 \ge c_2^{1/4}$ and (vii) implies (vi) if $c_2 \ge c_3$. □

Notice that properties (i) and (ii) above make sense even when $d = 0$. Therefore, we shall define a function $f : \mathbb{Z}_N \to D$ to be α-uniform of degree zero if

$\left|\sum_s f(s)\right|^2 \leq \alpha N^2$. Property (iv) now makes sense when $d = 1$. This definition will allow us to begin an inductive argument at an earlier and thus easier place.

The next result is the main one of this section. Although it will not be applied directly, it easily implies the results that are needed for later.

Theorem 3.2. *Let $k \geq 2$ and let $f_1, \dots, f_k$ be functions from $\mathbb{Z}_N$ to D such that f_k is α-uniform of degree $k-2$. Then*

$$\left|\sum_r \sum_s f_1(s) f_2(s-r) \dots f_k(s-(k-1)r)\right| \leq \alpha^{1/2^{k-1}} N^2 .$$

Proof. When $k = 2$, we know that

$$\left|\sum_r \sum_s f_1(s) f_2(s-r)\right| = \left|\Big(\sum_s f_1(s)\Big)\Big(\sum_t f_2(t)\Big)\right| \leq \alpha^{1/2} N^2 ,$$

since $\left|\sum_s f_1(s)\right| \leq N$ and $\left|\sum_t f_2(t)\right| \leq \alpha^{1/2} N$.

When $k > 2$, assume the result for $k-1$, let f_k be α-uniform of degree $k-2$ and let $\alpha : \mathbb{Z}_N \to [0,1]$ be a function with the property that $\Delta(f_k; r)$ is $\alpha(r)$-uniform of degree $k-3$ for every $r \in \mathbb{Z}_N$. Then

$$\begin{aligned}
\Big|\sum_r & \sum_s f_1(s) \dots f_k(s-(k-1)r)\Big|^2 \\
&\leq N \sum_s \Big|\sum_r f_1(s) f_2(s-r) \dots f_k(s-(k-1)r)\Big|^2 \\
&\leq N \sum_s \Big|\sum_r f_2(s-r) f_3(s-2r) \dots f_k(s-(k-1)r)\Big|^2 \\
&= N \sum_s \sum_r \sum_t f_2(s-r)\overline{f_2(s-t)} \dots f_k(s-(k-1)r)\overline{f_k(s-(k-1)t)} \\
&= N \sum_s \sum_r \sum_u f_2(s)\overline{f_2(s-u)} \dots f_k(s-(k-2)r)\overline{f_k(s-(k-2)r-(k-1)u)} \\
&= N \sum_s \sum_r \sum_u \Delta(f_2; u)(s) \Delta(f_3; 2u)(s-r) \dots \Delta(f_k; (k-1)u)(s-(k-2)r) .
\end{aligned}$$

Since $\Delta(f_k; (k-1)u)$ is $\alpha((k-1)u)$-uniform of degree $k-3$, our inductive hypothesis implies that this is at most $N \sum_u \alpha((k-1)u)^{1/2^{k-2}} N^2$, and since $\sum_u \alpha((k-1)u) \leq \alpha N$, this is at most $\alpha^{1/2^{k-2}} N^2$, which proves the result for k. □

The interest in Theorem 3.2 is of course that the expression on the left hand side can be used to count arithmetic progressions. Let us now define a set $A \subset \mathbb{Z}_N$ to be α-uniform of degree d if its balanced function is. (This definition makes sense when $d = 0$, but only because it applies to all sets.) The next result implies that a set A which is α-uniform of degree $d-2$ for some small α contains about the number of arithmetic progressions of length d that a random set of the same cardinality would have, where this means arithmetic progressions mod N. We shall then show how to obtain genuine progressions, which turns out to be a minor technicality, similar to the corresponding technicality in the proof of Roth's theorem.

Corollary 3.3. *Let $A_1, \dots, A_k$ be subsets of $\mathbb{Z}_N$, such that A_i has cardinality $\delta_i N$ for every i, and is $\alpha^{2^{i-1}}$-uniform of degree $i-2$ for every $i \geq 3$. Then*

$$\left|\sum_r |(A_1 + r) \cap \dots \cap (A_k + kr)| - \delta_1 \dots \delta_k N^2\right| \leq 2^k \alpha N^2 .$$

Proof. For each i, let f_i be the balanced function of A_i. Then

$$|(A_1 + r) \cap \cdots \cap (A_k + kr)| = \sum_s (\delta_1 + f_1(s - r)) \ldots (\delta_k + f_k(s - kr)) ,$$

so we can rewrite $|(A_1 + r) \cap \cdots \cap (A_k + kr)| - \delta_1 \ldots \delta_k N$ as

$$\sum_{B \subset [k], B \neq \emptyset} \prod_{i \notin B} \delta_i \sum_s \prod_{i \in B} f_i(s - ir) .$$

Now if $j = \max B$, then $\sum_r \sum_s \prod_{i \in B} f_i(s - ir)$ is at most $\alpha^{2^{j-1}/2^{j-1}} N^2$, by Theorem 3.2. It follows that

$$\begin{aligned} \Big| \sum_r |(A_1 + r) \cap \cdots \cap (A_k + kr)| - \delta_1 \ldots \delta_k N^2 \Big| &\leq \textstyle\sum_{B \subset [k], B \neq \emptyset} \prod_{i \notin B} \delta_i . \alpha N^2 \\ &= \alpha N^2 \Big(\textstyle\prod_{i=1}^k (1 + \delta_i) - 1 \Big) \end{aligned}$$

which is at most $2^k \alpha N^2$, as required. □

We now prove two simple technical lemmas.

Lemma 3.4. *Let $d \geq 1$ and let $f : \mathbb{Z}_N \to D$ be α-uniform of degree d. Then f is $\alpha^{1/2}$-uniform of degree $d - 1$.*

Proof. Our assumption is that

$$\sum_{a_1, \ldots, a_d} \Big| \sum_s \Delta(f; a_1, \ldots, a_d)(s) \Big|^2 \leq \alpha N^{k+2} .$$

By the Cauchy-Schwarz inequality, this implies that

$$\Big| \sum_{a_1, \ldots, a_d} \sum_s \Delta(f; a_1, \ldots, a_d)(s) \Big| \leq \alpha^{1/2} N^{k+1} ,$$

which, by the equivalence of properties (i) and (ii) in Lemma 3.1, proves the lemma. □

Lemma 3.5. *Let A be an α-uniform subset of $\mathbb{Z}_N$ of cardinality δN, and let P be an interval of the form $\{a + 1, \ldots, a + M\}$, where $M = \beta N$. Then $\big||A \cap P| - \beta\delta N\big| \leq \alpha^{1/4} N$.*

Proof. First, we can easily estimate the Fourier coefficients of the set P. Indeed,

$$\begin{aligned} |\tilde{P}(r)| &= \Big|\textstyle\sum_{s=1}^M \omega^{-r(a+s)}\Big| \\ &= |(1 - \omega^{rM})/(1 - \omega^r)| \leq N/2r . \end{aligned}$$

(We also know that it is at most M, but will not need to use this fact.) This estimate implies that $\sum_{r \neq 0} |\tilde{P}(r)|^{4/3} \leq N^{4/3}$. Therefore,

$$\big||A \cap P| - \beta\delta N\big| = N^{-1} \Big| \textstyle\sum_{r \neq 0} \hat{A}(r) \tilde{P}(r) \Big|$$

$$\leq N^{-1}\Big(\sum_{r\neq 0}|\hat{A}(r)|^4\Big)^{1/4}\Big(\sum_{r\neq 0}|\tilde{P}(r)|^{4/3}\Big)^{3/4}$$
$$\leq \Big(\sum_{r\neq 0}|\hat{A}(r)|^4\Big)^{1/4} \leq \alpha^{1/4}N$$

using property (iv) of Lemma 3.1.

Corollary 3.6. *Let $A \subset \mathbb{Z}_N$ be α-uniform of degree $k-2$ and have cardinality δN. If $\alpha \leq (\delta/2)^{k2^k}$ and $N \geq 32k^2\delta^{-k}$ then A contains an arithmetic progression of length k.*

Proof. Let $A_1 = A_2 = A \cap \big[(k-2)N/(2k-3), (k-1)N/(2k-3)\big]$, and let $A_3 = \cdots = A_k = A$. By Lemma 3.4 A is $\alpha^{1/2^{k-3}}$-uniform (of degree one), so by Lemma 3.5 the sets A_1 and A_2 both have cardinality at least $\delta N/4k$ since, by the first inequality we have assumed, we know that $\alpha^{1/2^{k-1}} \leq \delta/4k$.

Therefore, by Corollary 3.3, A contains at least $\big((\delta^k/16k^2) - 2^k\alpha^{1/2^{k-1}}\big)N^2$ arithmetic progressions modulo N with the first two terms belonging to the interval $\big[(k-2)N/(2k-3), (k-1)N/(2k-3)\big]$. The only way such a progression can fail to be genuine is if the common difference is zero, and there are at most δN such degenerate progressions. Thus the corollary is proved, since the two inequalities we have assumed imply that $(\delta^k/16k^2) - 2^k\alpha^{1/2^{k-1}} \geq \delta^k/32k^2$ and $\delta^k N^2/32k^2 > \delta N$. □

Remark. Notice that the proof of Corollary 3.6 did not use Fourier coefficients. This shows that in the proof of Theorem 2.6, the Fourier analysis was not really needed for the analysis of case 3. However, it was used in a more essential way for case 2.

In order to prove Szemerédi's theorem, it is now enough to prove that if $A \subset \mathbb{Z}_N$ is a set of size δN which is not $(\delta/2)^{k2^k}$-uniform of degree $d-2$, then there is an arithmetic progression $P \subset \mathbb{Z}_N$ of length tending to infinity with N, such that $|A \cap P| \geq (\delta+\epsilon)|P|$, where $\epsilon > 0$ depends on δ and d only. Thus, we wish to deduce a structural property of A from information about its differences. We do not quite have an inverse problem, as usually defined, of additive number theory, but it is certainly in the same spirit, and we shall relate it to a well-known inverse problem, Freiman's theorem, later in the paper. For the rest of this section we shall give a combinatorial characterization of α-uniform sets of degree d. The result will not be needed for Szemerédi's theorem but gives a little more insight into what is being proved. Also, Lemma 3.7 below will be used near the end of the paper.

Let A be a subset of $\mathbb{Z}_N$ and let $d \geq 0$. By a *d-dimensional cube in A* we shall mean a function $\phi : \{0,1\}^d \to A$ of the form

$$\phi : (\epsilon_1, \ldots, \epsilon_d) \mapsto a_0 + \epsilon_1 a_1 + \cdots + \epsilon_d a_d \ ,$$

where $a_0, a_1, \ldots, a_d$ all belong to $\mathbb{Z}_N$. We shall say that such a cube is contained in A, even though it is strictly speaking contained in $A^{\{0,1\}^d}$.

Let $A \subset \mathbb{Z}_N$ have cardinality δN. Then A obviously contains exactly δN cubes of dimension zero and $\delta^2 N^2$ cubes of dimension one. As remarked after Lemma

2.2, the number of two-dimensional cubes in A can be written as $N^{-1}\sum_r |\hat{A}(r)|^4$, so A is α-uniform if and only if there are at most $(\delta^4+\alpha)N^3$ of them. We shall now show that A contains at least $\delta^{2^d}N^{d+1}$ cubes of dimension d, and that equality is nearly attained if A is α-uniform of degree $d-1$ for some small α. The remarks we have just made prove this result for $d=1$. Notice that equality is also nearly attained (with high probability) if A is a random set of cardinality δN. This is why we regard higher-degree uniformity as a form of pseudorandomness.

Lemma 3.7. *Let A be a subset of $\mathbb{Z}_N$ of cardinality δN and let $d \geq 0$. Then A contains at least $\delta^{2^d}N^{d+1}$ cubes of dimension d.*

Proof. We know the result for $d=0$ or 1 so let $d>1$ and assume that the result is known for $d-1$. The number of d-dimensional cubes in A is the sum over all r of the number of $(d-1)$-dimensional cubes in $A\cap(A+r)$. Write $\delta(r)N$ for the cardinality of $A\cap(A+r)$. Then by induction the number of d-dimensional cubes in A is at least $\sum_r \delta(r)^{2^{d-1}}N^d$. Since the average value of $\delta(r)$ is exactly δ^2, this is at least $\delta^{2^d}N^{d+1}$ as required. □

The next lemma is little more than the Cauchy-Schwarz inequality and some notation. It will be convenient to use abbreviations such as x for $(x_1,\dots,x_k)$ and $x.y$ for $\sum_{i=1}^k x_iy_i$. If $\epsilon\in\{0,1\}^k$ then we shall write $|\epsilon|$ for $\sum_{i=1}^k\epsilon_i$. Once again, C is the operation of complex conjugation.

Lemma 3.8. *For every $\epsilon\in\{0,1\}^k$ let f_ϵ be a function from $\mathbb{Z}_N$ to D. Then*

$$\Big|\sum_{x\in\mathbb{Z}_N^d}\sum_s\prod_{\epsilon\in\{0,1\}^d}C^{|\epsilon|}f_\epsilon(s-\epsilon.x)\Big| \leq \prod_{\epsilon\in\{0,1\}^d}\Big|\sum_{x\in\mathbb{Z}_N^d}\sum_s\prod_{\eta\in\{0,1\}^d}C^{|\eta|}f_\epsilon(s-\eta.x)\Big|^{1/2^d}.$$

Proof.

$$\begin{aligned}&\Big|\textstyle\sum_{x\in\mathbb{Z}_N^d}\sum_s\prod_{\epsilon\in\{0,1\}^d}C^{|\epsilon|}f_\epsilon(s-\epsilon.x)\Big|\\ =&\Big|\textstyle\sum_{x\in\mathbb{Z}_N^{d-1}}\Big(\sum_s\prod_{\epsilon\in\{0,1\}^{d-1}}C^{|\epsilon|}f_{\epsilon,0}(s-\epsilon.x)\Big)\Big(\sum_t\prod_{\epsilon\in\{0,1\}^{d-1}}C^{|\epsilon|}f_{\epsilon,1}(t-\epsilon.x)\Big)\Big|\\ \leq&\Big(\textstyle\sum_{x\in\mathbb{Z}_N^{d-1}}\Big|\sum_s\prod_{\epsilon\in\{0,1\}^{d-1}}C^{|\epsilon|}f_{\epsilon,0}(s-\epsilon.x)\Big|^2\Big)^{\frac12}\\ &\Big(\textstyle\sum_{x\in\mathbb{Z}_N^d}\Big|\sum_s\prod_{\epsilon\in\{0,1\}^{d-1}}C^{|\epsilon|}f_{\epsilon,1}(s-\epsilon.x)\Big|^2\Big)^{\frac12}\end{aligned}$$

Let us write $P_d(\epsilon)$ and $Q_d(\epsilon)$ for the sequences $(\epsilon_1,\dots,\epsilon_{d-1},0)$ and $(\epsilon_1,\dots,\epsilon_{d-1},1)$. Then

$$\sum_{x\in\mathbb{Z}_N^{d-1}}\Big|\sum_s\prod_{\epsilon\in\{0,1\}^{d-1}}C^{|\epsilon|}f_{\epsilon,0}(s-\epsilon.x)\Big|^2=\sum_{x\in\mathbb{Z}_N^d}\sum_s\prod_{\epsilon\in\{0,1\}^d}C^{|\epsilon|}f_{P_d(\epsilon)}(s-\epsilon.x)$$

and similarly for the second bracket with Q_d, so the two parts are square roots of expressions of the form we started with, except that the function f_ϵ no longer

depends on ϵ_d. Repeating this argument for the other coordinates, we obtain the result. □

If we regard Lemma 3.8 as a modification of the Cauchy-Schwarz inequality, then the next lemma is the corresponding modification of Minkowski's inequality.

Lemma 3.9. *Given any function $f : \mathbb{Z}_N \to \mathbb{C}$ and any $d \geq 2$, define $\|f\|_d$ by the formula*

$$\|f\|_d = \Big| \sum_{x \in \mathbb{Z}_N^d} \sum_s \prod_{\epsilon \in \{0,1\}^d} C^{|\epsilon|} f(s - \epsilon.x) \Big|^{1/2^d} .$$

Then $\|f+g\|_d \leq \|f\|_d + \|g\|_d$ for any pair of functions $f, g : \mathbb{Z}_N \to \mathbb{C}$. In other words, $\|.\|_d$ is a norm.

Proof. If we expand $\|f+g\|^{2^d}$, we obtain the sum

$$\sum_{x \in \mathbb{Z}_N^d} \sum_s \prod_{\epsilon \in \{0,1\}^d} C^{|\epsilon|} (f+g)(s - \epsilon.x) .$$

If we expand the product we obtain 2^{2^d} terms of the form $\prod_{\epsilon \in \{0,1\}^d} C^{|\epsilon|} f_\epsilon(s - \epsilon.x)$, where each function f_ϵ is either f or g. For each one of these terms, if we take the sum over $x_1, \dots, x_d$ and s and apply Lemma 3.8, we have an upper estimate of $\|f\|_d^k \|g\|_d^l$, where k and l are the number of times that f_ϵ equals f and g respectively. From this it follows that

$$\|f+g\|^{2^d} \leq \sum_{k=0}^{2^d} \binom{2^d}{k} \|f\|_d^k \|g\|_d^{2^d-k} = (\|f\|_d + \|g\|_d)^{2^d}$$

which proves the lemma. □

It is now very easy to show that equality is almost attained in Lemma 3.7 for sets that are sufficiently uniform.

Lemma 3.10. *Let A be α-uniform of degree $d-1$. Then A contains at most $(\delta + \alpha^{1/2^d})^{2^d} N^{d+1}$ cubes of dimension d.*

Proof. Write $A = \delta + f$, where $|A| = \delta N$ and f is the balanced function of A. Then $\|A\|_d \leq \|\delta\|_d + \|f\|_d$. It is easy to see that $\|A\|_d^{2^d}$ is the number of d-dimensional cubes in A and that $\|\delta\|_d^{2^d} = \delta^{2^d} N^{d+1}$. Moreover, the statement that A is α-uniform of degree $d-1$ is equivalent to the statement that $\|f\|_d^{2^d} \leq \alpha N^{d+1}$. Therefore, Lemma 3.9 tells us that A contains at most $(\delta + \alpha^{1/2^d})^{2^d} N^{d+1}$ cubes of dimension d. □

Remark. In a sense, the normed spaces just defined encapsulate all the information we need about the arithmetical properties of the functions we consider. In their definitions they bear some resemblance to Sobolev spaces. Although I cannot think

of any potential applications, I still feel that it would be interesting to investigate them further.

4. Some Motivating Examples.

We now know that Szemerédi's theorem would follow from an adequate understanding of higher-degree uniformity. A natural question to ask is whether degree-one uniformity *implies* higher-degree uniformity (for which it would be enough to show that it implied quadratic uniformity). To make the question precise, if A has density δ and is α-uniform, does it follow that A is quadratically β-uniform, for some β depending on α and δ only? If so, then the same result for higher-degree uniformity can be deduced, and Szemerédi's theorem follows easily, by the method of §2.

I have already said that α-uniformity is not an appropriate definition of pseudorandomness for progressions of length greater than three, so the answer to this question is no. The first result of this section is a simple counterexample demonstrating this. Let A be the set $\{s \in \mathbb{Z}_N : |s^2| \leq N/100\}$. If $s \in A \cap (A+k)$, then $|s^2| \leq N/100$ and $|(s-k)^2| \leq N/100$ as well, which implies that $|2sk - k^2| \leq N/50$, or equivalently that s lies inside the set $(2k)^{-1}\{s : |s - k/2| \leq N/50\}$. It follows that $A \cap (A+k)$ is not uniform for any $k \neq 0$.

Let us also think about the number of mod-N arithmetic progressions of length four contained in A. Such a progression can be written in the form $(x-d, x, x+d, x+2d)$. Now $s \in A$ if and only if s^2 is 'small', in the sense that s^2 is close to zero. Suppose we know that $(x-d)^2$, x^2 and $(x+d)^2$ are all small in this sense. Then, by taking differences, we know that $2xd - d^2$ and $2xd + d^2$ are also small, from which it follows that $4xd$ and $2d^2$ are both small, from which it follows that $x^2 + 4xd + 4d^2 = (x+2d)^2$ is small. This means that it is more likely than it should be that $x + 2d$ also belongs to A, and therefore that A contains *more* arithmetic progressions of length four than would be contained in a random set of the same cardinality. (I am certain, but have not actually checked, that a similar example could be constructed with too few such progressions.)

It is possible, but not completely straightforward, to show that A itself is highly uniform. Rather than go into the details, we prove a closely related fact which is in some ways more natural. Let $f(s) = \omega^{s^2}$. We shall show that f is a very uniform function, while $\Delta(f;k)$ fails badly to be uniform for any $k \neq 0$. For the uniformity of f, notice that

$$|\hat{f}(r)| = \left|\sum_s \omega^{s^2 - rs}\right| = \left|\sum_s \omega^{(s-r/2)^2}\right| = \left|\sum_s \omega^{s^2}\right|$$

for every r. Therefore, $|\hat{f}(r)| = N^{1/2}$ for every $r \in \mathbb{Z}_N$, so f is as uniform as a function into the unit circle can possibly be. On the other hand, $\Delta(f;k)(s) = \omega^{2ks-k^2}$, so that

$$\Delta(f;k)\hat{}(r) = \begin{cases} N & \text{r=2k} \\ 0 & \textit{otherwise} \end{cases}$$

which shows that $\Delta(f;k)$ is, for $k \neq 0$, as non-uniform as possible.

More generally, if $\phi : \mathbb{Z}_N \to \mathbb{Z}_N$ is a quadratic polynomial and $f(s) = \omega^{\phi(s)}$, then f is highly uniform, but there is some $\lambda \in \mathbb{Z}_N$ such that, for every k,

$$\Delta(f;k)\hat{}(r) = \begin{cases} N & r = \lambda k \\ 0 & otherwise. \end{cases}$$

This suggests an attractive conjecture, which could perhaps replace the false idea that if A is uniform then so are almost all $A \cap (A+k)$. Perhaps if there are many values of k for which $A \cap (A+k)$ fails to be uniform, then there must be a quadratic function $\phi : \mathbb{Z}_N \to \mathbb{Z}_N$ such that $\left|\sum_{s\in A} \omega^{-\phi(s)}\right|$ is large. We shall see in the next section that such "quadratic bias" would actually imply the existence of a long arithmetic progression P_j such that $|A \cap P_j|/|P_j|$ was significantly larger than $|A|/N$. This would give a proof of Szemerédi's theorem for progressions of length four, and one can see how the above ideas might be generalized to higher-degree polynomials and longer arithmetic progressions.

The second example of this section shows that such conjectures are still too optimistic. As with the first example, we shall consider functions that are more general than characteristic functions of subsets of $\mathbb{Z}_N$. However, this should be enough to convince the reader not to try to prove the conjectures.

Let r be about $\sqrt{N}$ and for $0 \leq a, b < r/2$ define $\phi(ar+b)$ to be $a^2 + b^2$. Now define

$$f(s) = \begin{cases} \omega^{\phi(s)} & s = ar + b for some 0 \leq a, b < r/2 \\ 0 & otherwise. \end{cases}$$

The function f is not quadratic, but it resembles a quadratic form in two variables (with the numbers 1 and r behaving like a basis of a two-dimensional space).

Suppose $s = ar + b$ and $k = cr + d$ are two numbers in $\mathbb{Z}_N$, where all of a, b, $a - c$ and $b - d$ lie in the interval $[0, r/2)$. Then

$$f(s)\overline{f(s-k)} = \omega^{2ac - c^2 + 2bd - d^2} = \omega^{\phi_k(ar+b)}$$

where ϕ_k depends linearly on the pair (a, b). The property that will interest us about ϕ_k is that, at least when c and d are not too close to $r/2$, there are several pairs (a, b) such that the condition on (a, b, c, d) applies, and therefore several quadruples $\big((a_i, b_i)\big)_{i=1}^4$ such that

$$(a_1, b_1) + (a_2, b_2) = (a_3, b_3) + (a_4, b_4)$$

and

$$\phi_k(a_1 r + b_1) + \phi_k(a_2 r + b_2) = \phi_k(a_3 r + b_3) + \phi_k(a_4 r + b_4) \ .$$

Here, "several" means a number proportional to N^3, which is the maximum it could be.

Let B be the set of all $s = ar + b$ for which a, b, c and d satisfy the conditions above. (Of course, B depends on k.) Then

$$\sum_q \Big|\sum_{s\in B} \omega^{\phi_k(s) - qs}\Big|^4 = N \sum \{\omega^{\phi_k(s)+\phi_k(t)-\phi_k(u)-\phi_k(v)} : s, t, u, v \in B,\ s+t = u+v\} \ .$$

Now the set B has been chosen so that if $s, t, u, v \in B$ and $s + t = u + v$, then $\phi_k(s) + \phi_k(t) = \phi_k(u) + \phi_k(v)$. Therefore, the right hand side above is N times the number of quadruples $(s, t, u, v) \in B^4$ such that $s + t = u + v$. It is not hard to check that if c and d are smaller than $r/4$, say, then B has cardinality proportional to N^3, and therefore that the right hand side above is proportional to N^4. Lemma

2.1 now tells us that ϕ_k has a large Fourier coefficient. Thus, at the very least, we have shown that, for many values of k, $\Delta(f;k)$ fails to be uniform.

If we could find a *genuinely* quadratic function $\phi(s) = as^2 + bs + c$ such that $\left|\sum_s f(s)\omega^{-\phi(s)}\right|^2$ was proportional to N^2, then, expanding, we would have

$$\sum_{s,k} f(s)\overline{f(s-k)}\omega^{-\phi(s)+\phi(s-k)} = \sum_{s,k} f(s)\overline{f(s-k)}\omega^{-2ask-bk}$$

proportional to N^2, which would imply that the number of k for which $\Delta(f;k)\hat{}(2ak)$ was proportional to N was proportional to N. A direct calculation (left to the interested reader) shows that such a phenomenon does not occur. That is, there is no value of λ such that $\Delta(f;k)\hat{}(\lambda k)$ is large for many values of k.

There are of course many examples like the second one above. One can define functions that resemble d-dimensional quadratic forms, and provided that d is small the same sort of behaviour occurs. Thus, we must accept that the ideas of this paper so far do not lead directly to a proof of Szemerédi's theorem, and begin to come to terms with these "multi-dimensional" examples. It is for this purpose that our major tool, an adaptation of Freiman's theorem, is used, as will be explained later in the paper.

5. Consequences of Weyl's inequality.

In the proof of Roth's theorem an important role was played by Lemma 2.3, which said that if $\phi : \{1,2,\dots,r\} \to \mathbb{Z}_N$ is a linear function (in the sense that it can be defined by a formula $\phi(x) = ax + b$) then the set $\{1,2,\dots,r\}$ can be partitioned into not too many arithmetic progressions, on each of which ϕ is roughly constant. The linear function arose because we were using the hypothesis that A had a large Fourier coefficient - that is, that $\sum_x A(x)\omega^{-rx}$ was large for some r. The linear function in question was then $x \mapsto rx$.

In a certain sense, then, a set that failed to be α-uniform was *linearly biased.* We shall see later that a set that fails to be α-quadratically uniform is *quadratically biased.* In a tidy world, this would mean that there was a quadratic function q with the property that $\sum_x A(x)\omega^{-q(x)}$ was large. Actually, all we can show (necessarily, in the light of the second example of the previous section) is that $\mathbb{Z}_N$ can be partitioned into a smallish number of subprogressions P_i in each of which there is a function q_i such that $\left|\sum_{x\in P_i}\omega^{-q_i(x)}\right|$ is, on average, an appreciable fraction of $|P|$.

This result explains our use of the term 'quadratically uniform'. A set that fails to be quadratically uniform exhibits a certain sort of quadratic bias. For this to be useful to us, we need to show that, given such a biased set, we can pass to a subprogression inside which it has increased density. Then, just as in Roth's theorem, we will be able to iterate. Rather than give the details of this step, I shall sketch the argument.

The main lemma of this part of the proof states that if $q : \{1,2,\dots,r\} \to \mathbb{Z}$ is a quadratic function, then $\{1,2,\dots,r\}$ can be partitioned into subprogressions P_i on each of which q is roughly constant. The average size of the P_i turns out to be r^c for some absolute constant $c > 0$.

An interesting result of Furstenberg states that the sequence of squares is recurrent, in the sense that, given any measure-preserving dynamical system (X, μ, T) and any set A of positive measure in X, there must exist a positive integer n such that $A \cap T^{-n^2}A$ has positive measure. (There is a similar statement for topological dynamical systems: every point x must return arbitrarily close to its starting point after a square number of iterations.) From this it is easy to deduce that, given any real number α and any $\epsilon > 0$ there exists a positive integer n such that $n^2\alpha$ is within ϵ of an integer. Equivalently, if one wishes to approximate α by a fraction of the form p/q^2, then one can do so, and can beat the trivial bound of $|\alpha - p/q^2| \leq c/q^2$.

This corollary was in fact first proved by Weyl, and from Weyl's argument it follows that one can find p and q with $|\alpha - p/q^2| \leq c/q^{2+\gamma}$ for some fixed $\gamma > 0$. Let us first see roughly how Weyl's argument works and then how the result is used for the purpose of partitioning the progression $\{1, 2, \dots, r\}$ above.

The first step is closely related to Weyl's equidistribution theorem. If A is a random subset of $\mathbb{Z}_N$ of cardinality r, then the expected size of the intersection $A \cap \{-M, -(M-1), \dots, (M-1), M\}$ is roughly $2Mr/N$. If rM is significantly larger than N and A has *empty* intersection with $\{-M, -(M-1), \dots, (M-1), M\}$, then A is in some sense not random. Moreover, it fails to be random in a way that ought to be detected by the Fourier coefficients of the characteristic function of A, since in a certain sense A is not evenly distributed round the circle.

Here is a precise result along these lines.

Lemma 5.1. *Let A be a subset of $\mathbb{Z}_N$ of cardinality r, let M be an even integer and suppose that $A \cap [-M, M) = \emptyset$. Then there exists u with $0 < |u| \leq N^2M^{-2}$ such that $|\hat{A}(u)| \geq kM/2N$.*

If we apply this lemma to the set $\{q(1), q(2), \dots, q(r)\}$ and assume that for no value of $x \leq r$ does $q(x)$ lie in the interval $\{-M, -(M-1), \dots, (M-1), M\}$ for some suitably chosen M (significantly larger than N/r but significantly smaller than N) then we discover from Lemma 5.1 that there is a small u such that $\sum_{x=1}^{r} \omega^{uq(x)}$ is large.

If we write α for the number u/N, then this sum can be rewritten $\sum_{x=1}^{r} e(\alpha q(x))$, where $e(t)$ stands for $\exp(2\pi it)$. A great deal is known about such sums, and one fundamental result, due to Weyl (it has since been surpassed but is sufficient for our purposes) is that such a sum can be large only if α is close to a rational number with a small denominator. (To see that some condition like this is necessary, one has only to consider a sum like $\sum_{x=1}^{r} e(x^2/3)$. The summand equals 1 a third of the time and $\exp(2\pi i/3)$ for the other two thirds, and therefore the sum has magnitude proportional to r, which counts as very large indeed since it is close to the trivial upper bound of r.) For a detailed statement and proof, see [V]. The result originally appeared in [We].

Now consider a function from $\{1, 2, \dots, r\}$ to $\mathbb{Z}_N$ of the form $x \mapsto ax^2$. We know that if a/N is not close to a rational of small denominator then the image of this function is well distributed round the circle, and in particular contains values close to zero mod N. On the other hand, if a/N *is* close to a rational p/q with q small, then aq^2/N is close to the integer pq, from which it follows that aq^2 is close to pqN, and therefore that aq^2 is close to zero mod N. In other words, in this case the result holds for a very simple reason. Needless to say, one has to check that the numbers work out in the above argument, but, happily, they do.

Now let $q : \{1, 2, \dots, r\} \to \mathbb{Z}_N$ be defined by the formula $x \mapsto ax^2 + bx$ and let us try to partition $\{1, 2, \dots, r\}$ into subprogressions on which q is almost constant. (There is no loss of generality in assuming that q has a zero constant term.) We begin by finding some t significantly smaller than r such that at^2 is close to 0. (This we do by applying our result above to a small subinterval $\{1, 2, \dots, s\}$, say.) If at^2 is close to 0, then $a(x + ht)^2 - ax^2 = 2axht + ah^2t^2$ which, if h is small, is close to $2axht$. Since $2axht$ depends linearly on h, for fixed x, it follows that if we partition $\{1, 2, \dots, r\}$ into sufficiently short arithmetic progressions, each with common difference t, the function $x \mapsto ax^2$ is approximately linear on each of these progressions (though the 'gradients' will differ from progression to progression).

This reduces the problem to one we already know how to solve: given a linear function from an arithmetic progression to $\mathbb{Z}_N$, partition that progression into smaller ones on each of which the function is approximately constant. We proved exactly such a result in Lemma 2.3.

By similar techniques, one can prove more or less any result of this kind that one might need. For the general case of Szemerédi's theorem, it is necessary to prove generalizations to polynomials of higher degree, and several of them simultaneously. Another result that is needed is that if P is a progression and $\mu : P \to \mathbb{Z}_N$ is defined multilinearly, then P^d can be partitioned into subsets of the form $Q_1 \times \cdots \times Q_d$, where the Q_i are subprogressions of P with the same common difference and of roughly the same size, on each of which μ is roughly constant. And in fact this is needed for several multilinear functions simultaneously.

6. Somewhat Additive Functions.

We saw in §4 that it is possible for a set A to have small Fourier coefficients, but for $A \cap (A + k)$ to have at least one non-trivial large Fourier coefficient for every k. Moreover, the obvious conjecture concerning such sets, that they correlate with some function of the kind $\omega^{q(s)}$ where q is a quadratic polynomial, is false. The aim of the next three sections is to show that such a set A must nevertheless exhibit quadratic bias of some sort. We will then be able to use the results of the last section to find linear bias, which will complete the proof for progressions of length four. The generalization to longer progressions will use similar ideas, but involves one extra important difficulty.

Notice that what we are trying to prove is very natural. If we replace A by a function on $\mathbb{Z}_N$ of the form $f(s) = \omega^{\phi(s)}$, where $\phi : \mathbb{Z}_N \to \mathbb{Z}_N$, then we are trying to prove that if, for many k, the function $\phi_k(s) = \phi(s) - \phi(s - k)$ has some sort of linearity property, resulting in a large Fourier coefficient for the difference function $\Delta(f; k) = \omega^{\phi(s)-\phi(s-k)}$, then ϕ itself must in some way be quadratic. Many arguments in additive number theory (in particular Weyl's inequality) use the fact that taking difference functions reduces the degree of, and hence simplifies, a polynomial. We are trying to do something like the reverse process, "integrating" rather than "differentiating" and showing that the degree goes up by one. This is another sense in which we are engaged in an inverse problem.

This section contains a simple but crucial observation, which greatly restricts the possibilities for the Fourier coefficients of $A \cap (A + k)$ that are large. Let A be a set which is not quadratically α-uniform and let f be the balanced function of A.

Then by the definition of quadratic α-uniformity there are at least αN values of k such that we can find r for which

$$\Big|\sum_s f(s)f(s-k)\omega^{-rs}\Big| \geq \alpha N .$$

Letting B be the set of k for which such an r exists, we can find a function $\phi : B \to \mathbb{Z}_N$ such that

$$\sum_{k\in B}\Big|\sum_s f(s)f(s-k)\omega^{-\phi(k)s}\Big|^2 \geq \alpha^3 N^3 .$$

We shall show that the function ϕ has a weak-seeming property which we shall call γ-additivity, for a certain constant $\gamma > 0$ to be defined later. Using a variant of Freiman's theorem proved in the next section, we shall show that this property gives surprisingly precise information about ϕ.

Proposition 6.1. *Let $\alpha > 0$, let $f : \mathbb{Z}_N \to D$, let $B \subset \mathbb{Z}_N$ and let $\phi : B \to \mathbb{Z}_N$ be a function such that*

$$\sum_{k\in B}\big|\Delta(f;k)\hat{}\,(\phi(k))\big|^2 \geq \alpha N^3 .$$

Then there are at least $\alpha^4 N^3$ quadruples $(a,b,c,d) \in B^4$ such that $a+b = c+d$ and $\phi(a)+\phi(b) = \phi(c)+\phi(d)$.

Proof. Expanding the left hand side of the inequality we are assuming gives us the inequality

$$\sum_{k\in B}\sum_{s,t} f(s)\overline{f(s-k)f(t)}f(t-k)\omega^{-\phi(k)(s-t)} \geq \alpha N^3 .$$

If we now introduce the variable $u = s-t$ we can rewrite this as

$$\sum_k\sum_{s,u} f(s)\overline{f(s-k)f(s-u)}f(s-k-u)\omega^{-\phi(k)u} \geq \alpha N^3 .$$

Since $|f(x)| \leq 1$ for every x, it follows that

$$\sum_u\sum_s\Big|\sum_{k\in B}\overline{f(s-k)}f(s-k-u)\omega^{-\phi(k)u}\Big| \geq \alpha N^3$$

which implies that

$$\sum_u\sum_s\Big|\sum_{k\in B}\overline{f(s-k)}f(s-k-u)\omega^{-\phi(k)u}\Big|^2 \geq \alpha^2 N^4 .$$

For each u and x let $f_u(x) = \overline{f(-x)}f(-x-u)$ and let $g_u(x) = B(x)\omega^{\phi(x)u}$. The above inequality can be rewritten

$$\sum_u\sum_s\Big|\sum_k f(k-s)\overline{g(k)}\Big|^2 \geq \alpha^2 N^4 .$$

By Lemma 2.1, we can rewrite it again as

$$\sum_u\sum_r |\hat{f}_u(r)|^2|\hat{g}_u(r)|^2 \geq \alpha^2 N^5 .$$

Since $\sum_r |\hat{f}(r)|^4 \leq N^4$, the Cauchy-Schwarz inequality now implies that

$$\sum_u \Big(\sum_r |\hat{g}_u(r)|^4\Big)^{1/2} \geq \alpha^2 N^3 .$$

Applying the Cauchy-Schwarz inequality again, we can deduce that

$$\sum_{u,r} |\hat{g}(r)|^4 = \sum_{u,r} \Big|\sum_{k\in B} \omega^{\phi(s)u-rs}\Big|^4 \geq \alpha^4 N^5 .$$

Expanding the left hand side of this inequality we find that

$$\sum_{u,r} \sum_{a,b,c,d\in B} \omega^{u(\phi(a)+\phi(b)-\phi(c)-\phi(d))}\omega^{-r(a+b-c-d)} \geq \alpha^4 N^5 .$$

But now the left hand side is exactly N^2 times the number of quadruples $(a,b,c,d) \in B^4$ for which $a+b=c+d$ and $\phi(a)+\phi(b) = \phi(c)+\phi(d)$. This proves the proposition. □

If G is an Abelian group and a,b,c,d are elements of G such that $a+b=c+d$, we shall say that (a,b,c,d) is an *additive quadruple*. Given a subset $B \subset \mathbb{Z}_N$ and a function $\phi : B \to \mathbb{Z}_N$, let us say that a quadruple $(a,b,c,d) \in B^4$ is *ϕ-additive* if it is additive and in addition $\phi(a)+\phi(b) = \phi(c)+\phi(d)$. Let us say also that ϕ is *γ-additive* if there are at least γN^3 ϕ-additive quadruples. It is an easy exercise to show that if $\gamma = 1$ then B must be the whole of $\mathbb{Z}_N$ and $\phi : \mathbb{Z}_N \to \mathbb{Z}_N$ must be of the form $\phi(x) = \lambda x + \mu$, i.e., linear. Notice that the property of γ-additivity appeared, undefined, in §4 during the discussion of the function ϕ_k. Let us now give a simple but useful reformulation of the concept of γ-additivity.

Lemma 6.2. *Let $\gamma > 0$, let $B \subset \mathbb{Z}_N$, let $\phi : B \to \mathbb{Z}_N$ be a γ-additive function and let $\Gamma \subset \mathbb{Z}_N^2$ be the graph of ϕ. Then Γ contains at least γN^3 additive quadruples (in the group $\mathbb{Z}_N^2$).* □

As we have just remarked, a 1-additive function must be a linear. We finish this section with an important (and, in the light of the second example of §4, natural) example of a γ-additive function which cannot be approximated by a linear function even though γ is reasonably large. Let $x_1,\dots,x_d \in Z_N$ and $r_1,\dots,r_d \in \mathbb{N}$ be such that all the numbers $\sum_{i=1}^d a_i x_i$ with $0 \leq a_i < r_i$ are distinct. Let $y_1,\dots,y_d \in \mathbb{Z}_N$ be arbitrary, and define

$$\phi\Big(\sum_{i=1}^d a_i x_i\Big) = \sum_{i=1}^d a_i y_i .$$

Let $\phi(s)$ be arbitrary for the other values of s. Then a simple calculation shows that the number of additive quadruples is at least $(2/3)^d r_1^3 \dots r_d^3$. If $r_1 \dots r_d = \beta N$, then ϕ is $(2/3)^d\beta^3$-additive.

The function ϕ resembles a linear map between vector spaces, and the number d can be thought of as the dimension of the domain of the ϕ. In the next two sections we shall show that all γ-additive functions have, at least in part, something like the above form, with d not too large and $r_1 \dots r_d$ an appreciable fraction of N (both depending, of course, on γ).

7. Variations on a Theorem of Freiman.

The results of the previous section tell us that if f is a function from $\mathbb{Z}_N$ to the unit disc D, B is a large subset of $\mathbb{Z}_N$ and $\phi : B \to \mathbb{Z}_N$ is a function such that $\Delta(f;k)\hat{}(\phi(k))$ is large for every $k \in B$, then the graph Γ of ϕ contains many quadruples (x,y,z,w) such that $x+y=z+w$ (in $\mathbb{Z}_N^2$). The purpose of this section is to introduce a theorem of Freiman, and to explain how it helps to tell us about the structure of a function with such a graph.

It will not be possible to apply Freiman's theorem to Γ straight away, since Γ does not satisfy the necessary hypotheses. One of the main results of this section is that Γ has a large subset that does satisfy them. This result, a quantitative version of a theorem of Balog and Szemerédi, has recently been used by Bourgain, in a modified form, to obtain the best known lower bounds for the Hausdorff dimensions of Kakeya sets in $\mathbb{R}^n$ when n is large. Before we discuss it, let us see what Freiman's theorem says.

Let A be a subset of $\mathbb{Z}$ of cardinality m. It is easy to see that $A+A = \{x+y : x,y \in A\}$ has cardinality between $2m-1$ and $m(m+1)/2$. Suppose that $|A+A| \leq Cm$ for some constant C. What information does this give about the set A? This problem is called an *inverse* problem of additive number theory, since it involves deducing the structure of A from the behaviour of $A+A$ - in contrast to a *direct* problem where properties of A give information about $A+A$.

It is clear that $A+A$ will be small when A is a subset of an arithmetic progression of length not much greater than m. After a moment's thought, one realises that there are other examples. For instance, one can take a "progression of progressions" such as $\{aM+b : 0 \leq a < h, 0 \leq b < k\}$ where $M >> k$ and $hk = m$. This example can then be generalized to a large subset of a "d-dimensional" arithmetic progression, provided that d is reasonably small. Freiman's theorem is the beautiful result that these simple examples exhaust all possibilities [F1,2]. A precise statement of the theorem is as follows.

Theorem 7.1. *Let C be a constant. Then there exist constants d_0 and K depending only on C such that whenever A is a subset of $\mathbb{Z}$ with $|A| = m$ and $|A+A| \leq Cm$, there exist $d \leq d_0$, an integer x_0 and positive integers $x_1,\dots,x_d$ and $k_1,\dots,k_d$ such that $k_1k_2\dots k_d \leq Km$ and*

$$A \subset \left\{x_0 + \sum_{i=1}^{d} a_i x_i : 0 \leq a_i < k_i \ (i = 1,2,\dots,d)\right\}.$$

The same is true if $|A-A| \leq Cm$.

It is an easy exercise to deduce from Theorem 7.1 the same result for subsets of $\mathbb{Z}^n$, where $x_0, x_1, \dots, x_d$ are now points in $\mathbb{Z}^n$. We shall in fact be interested in the case $n=2$, since we shall be applying Freiman's theorem to a graph coming from Proposition 6.1 and Lemma 6.2.

The number $k_1k_2\dots k_d$ is called the *size* of the d-dimensional arithmetic progression. Note that this is not necessarily the same as the cardinality of the set since there may be numbers (or more generally points of $\mathbb{Z}^D$) which can be written in more than one way as $x_0 + \sum_{i=1}^{d} a_i x_i$. When every such representation is

unique, we shall call the set a *proper* d-dimensional arithmetic progression. (This terminology is all standard.)

Freiman's original proof of Theorem 7.1 was long and very difficult to understand. Although a simplified version of his argument now exists [Bi], an extremely important breakthrough came a few years ago with a new and much easier proof by Ruzsa, which also provided a reasonable bound. This improved bound is very important for the purposes of our bound for Szemerédi's theorem. Full details of Ruzsa's proof can be found in [Ru1,2,3] or in a book by Nathanson [N], which also contains all necessary background material.

We shall in fact need a modification of Freiman's theorem, in which the hypothesis and the conclusion are weakened. In its qualitative form, the modification is a result of Balog and Szemerédi. However, they use Szemerédi's uniformity lemma, which for us is too expensive. Our argument will avoid the use of the uniformity lemma and thereby produce a much better bound than the bound of Balog and Szemerédi. It will be convenient (though not essential) to consider the version of Freiman's theorem where $A - A$, rather than $A + A$ is assumed to be small. Our weaker hypothesis concerns another parameter associated with a set A, which has several descriptions, and which we know applies to the graph Γ discussed above. It is

$$\|A * A\|_2^2 = \sum_{k \in \mathbb{Z}} |A \cap (A + k)|^2 = |\{(a, b, c, d) \in A^4 : a - b = c - d\}| .$$

(Freiman calls this invariant M' in his book [F2 p.41].) It is a straightforward exercise to show that

$$\|A * A\|_2^2 \leq m^2 + 2(1^2 + \cdots + (m-1)^2)$$

with equality if and only if A is an arithmetic progression of length m. The Balog-Szemerédi theorem [BS] is the following result.

Theorem 7.2. *Let A be a subset of $\mathbb{Z}^D$ of cardinality m and suppose that $\|A * A\|_2^2 \geq c_0 m^3$. Then there are constants c, K and d_0 depending only on c_0 and an arithmetic progression P of dimension $d \leq d_0$ and size at most Km such that $|A \cap P| \geq cm$.*

This result states that if $\|A * A\|_2^2$ is, to within a constant, as big as possible, then A has a proportional subset satisfying the conclusion of Freiman's theorem. Notice that, qualitatively at least, the conclusion of Theorem 7.2 cannot be strengthened, since if A has a proportional subset B with $\|B * B\|_2^2$ large, then $\|A * A\|_2^2$ is large whatever $A \setminus B$ is. To see that the new hypothesis is weaker, notice that if $|A - A| \leq Cm$, then $A \cap (A + k)$ is empty except for at most Cm values of k, while $\sum_{k \in \mathbb{Z}} |A \cap (A + k)| = m^2$. It follows from the Cauchy-Schwarz inequality that $\sum_{k \in \mathbb{Z}} |A \cap (A + k)|^2 \geq m^3/C$.

The most obvious approach to deducing Theorem 7.2 from Theorem 7.1 is to show that a set satisfying the hypothesis of Theorem 7.2 has a large subset satisfying the hypothesis of Theorem 7.1. This is exactly what Balog and Szemerédi did and we shall do as well.

Proposition 7.3. *Let A be a subset of $\mathbb{Z}^n$ of cardinality m such that $\|A * A\|_2^2 \geq c_0 m^3$. There are constants c and C depending only on c_0 and a subset $A'' \subset A$ of*

cardinality at least cm such that $|A'' - A''| \leq Cm$. *Moreover, c and C can be taken as* $2^{-20}c_0^{12}$ *and* $2^{38}c_0^{-24}$ *respectively.*

We shall need the following lemma for the proof.

Lemma 7.4. *Let V be a set of size m, let* $\delta > 0$ *and let* $A_1, \dots, A_n$ *be subsets of V such that* $\sum_{x=1}^n \sum_{y=1}^n |A_x \cap A_y| \geq \delta^2 mn^2$. *Then there is a subset* $K \subset [n]$ *of cardinality at least* $2^{-1/2}\delta^5 n$ *such that for at least* 90% *of the pairs* $(x, y) \in K^2$ *the intersection* $A_x \cap A_y$ *has cardinality at least* $\delta^2 m/2$. *In particular, the result holds if* $|A_x| \geq \delta m$ *for every x.*

Proof. For every $j \leq m$ let $B_j = \{i : j \in A_i\}$ and let $E_j = B_j^2$. Choose five numbers $j_1, \dots, j_5 \leq m$ at random (uniformly and independently), and let $X = E_{j_1} \cap \dots \cap E_{j_5}$. The probability p_{xy} that a given pair $(x, y) \in [n]^2$ belongs to E_{j_r} is $m^{-1}|A_x \cap A_y|$, so the probability that it belongs to X is p_{xy}^5. By our assumption we have that $\sum_{x,y=1}^n p_{xy} \geq \delta^2 n^2$, which implies (by Hölder's inequality) that $\sum_{x,y=1}^n p_{xy}^5 \geq \delta^{10} n^2$. In other words, the expected size of X is at least $\delta^{10} n^2$.

Let Y be the set of pairs $(x, y) \in X$ such that $|A_x \cap A_y| < \delta^2 m/2$, or equivalently $p_{xy} < \delta^2/2$. Because of the bound on p_{xy}, the probability that $(x, y) \in Y$ is at most $(\delta^2/2)^5$, so the expected size of Y is at most $\delta^{10} n^2/32$.

It follows that the expectation of $|X| - 16|Y|$ is at least $\delta^{10} n^2/2$. Hence, there exist $j_1, \dots, j_5$ such that $|X| \geq 16|Y|$ and $|X| \geq \delta^{10} n^2/2$. It follows that the set $K = B_{j_1} \cap \dots \cap B_{j_5}$ satisfies the conclusion of the lemma. □

Proof of Proposition 7.3. The function $f(x) = A * A(x)$ (from $\mathbb{Z}^n$ to $\mathbb{Z}$)is non-negative and satisfies $\|f\|_\infty \leq m$, $\|f\|_2^2 \geq c_0 m^3$ and $\|f\|_1 = m^2$. This implies that $f(x) \geq c_0 m/2$ for at least $c_0 m/2$ values of x, since otherwise we could write $f = g + h$ with g and h disjointly supported, g supported on fewer than $c_0 m/2$ points and $\|h\|_\infty \leq c_0 m/2$, which would tell us that

$$\|f\|_2^2 \leq \|g\|_2^2 + \|h\|_\infty \|h\|_1 < (c_0 m/2)m^2 + (c_0 m/2).m^2 = c_0 m^3 .$$

Let us call a value of x for which $f(x) \geq c_0 m/2$ a *popular difference* and let us define a graph G with vertex set A by joining a to b if $b - a$ (and hence $a - b$) is a popular difference. The average degree in G is at least $c_0^2 m/4$, so there must be at least $c_0^2 m/8$ vertices of degree at least $c_0^2 m/8$. Let $\delta = c_0^2/8$, let $a_1, \dots, a_n$ be vertices of degree at least $c_0^2 m/8$, with $n \geq \delta m$, and let $A_1, \dots, A_n$ be the neighbourhoods of the vertices $a_1, \dots, a_n$. By Lemma 7.4 we can find a subset $A' \subset \{a_1, \dots, a_n\}$ of cardinality at least $\delta^5 n/\sqrt{2}$ such that at least 90% of the intersections $A_i \cap A_j$ with $a_i, a_j \in A'$ are of size at least $\delta^2 m/2$. Set $\alpha = \delta^6/\sqrt{2}$ so that $|A'| \geq \alpha m$.

Now define a graph H with vertex set A', joining a_i to a_j if and only if $|A_i \cap A_j| \geq \delta^2 m/2$. The average degree of the vertices in H is at least $(9/10)|A'|$, so at least $|A'|/2$ vertices have degree at least $4|A'|/5$. Define A'' to be the set of all such vertices.

We claim now that A'' has a small difference set. To see this, consider any two elements $a_i, a_j \in A''$. Since the degrees of a_i and a_j are at least $(4/5)|A'|$ in H, there are at least $(3/5)|A'|$ points $a_k \in A'$ joined to both a_i and a_j. For every such k we have $|A_i \cap A_k|$ and $|A_j \cap A_k|$ both of size at least $\delta^2 m/2$. If $b \in A_i \cap A_k$, then

both $a_i - b$ and $a_k - b$ are popular differences. It follows that there are at least $c_0^2 m^2/4$ ways of writing $a_i - a_k$ as $(p-q)-(r-s)$, where $p,q,r,s \in A$, $p-q = a_i - b$ and $r - s = a_k - b$. Summing over all $b \in A_i \cap A_k$, we find that there are at least $\delta^2 c_0^2 m^3/8$ ways of writing $a_i - a_k$ as $(p-q)-(r-s)$ with $p,q,r,s \in A$. The same is true of $a_j - a_k$. Finally, summing over all k such that a_k is joined in H to both a_i and a_j, we find that there are at least $(3/5)|A'|\delta^4 c_0^4 m^6/64 \geq \alpha\delta^4 c_0^4 m^7/120$ ways of writing $a_i - a_j$ in the form $(p-q)-(r-s)-\big((t-u)-(v-w)\big)$ with $p,q,\ldots,w \in A$.

Since there are at most m^8 elements in A^8, the number of differences of elements of A'' is at most $120m/\alpha\delta^4 c_0^4 \leq 2^{38} m/c_0^{24}$. Note also that the cardinality of A'' is at least $(1/2)\alpha m \geq c_0^{12} m/2^{20}$. The proposition is proved. □

It is possible to apply Theorem 7.2 as it stands in order to prove Szemerédi's theorem for progressions of length four (and quite possibly in general). However, it is better to combine Proposition 7.3 with a weaker version of Freiman's theorem that gives less information about the structure of a set A with small difference set. There are three advantages in doing this. The first is that with the weaker version one can get a much better bound. The second is that using the weaker version is cleaner, particularly when we come to the general case. The third is that the weaker version is easier to prove than Freiman's theorem itself, as it avoids certain arguments from the geometry of numbers.

Rather than give the details, let me state a precise result and then explain in broad terms why Theorem 7.2 should lead us to expect such a result. I have given explicit constants just to stress that the result is completely effective, but some readers might prefer to replace expressions like $N^{2^{-3770}\gamma^{2328}\alpha^2}$ by ones like $N^{c\gamma^K\alpha^2}$.

Corollary 7.5. *Let N be sufficiently large, let $B_0 \subset \mathbb{Z}_N$ have cardinality αN and let $\phi : B_0 \to \mathbb{Z}_N$ have $\gamma(\alpha N)^3$ additive quadruples. Then there exist a mod-N arithmetic progression P of length at least $N^{2^{-3770}\gamma^{2328}\alpha^2}$, a subset $H \subset P$ of cardinality at least $2^{-1849}\gamma^{1164}\alpha|P|$ and constants $\lambda, \mu \in \mathbb{Z}_N$ such that $\phi(s) = \lambda s + \mu$ for every $s \in H$.*

What this corollary says is that if there are many ϕ-additive quadruples, then there must be a reasonably long arithmetic progression P and a linear function ψ such that $\phi(x) = \psi(x)$ for many values of $x \in P$. (Here, as in several other places, the word 'many' means 'as many as possible, up to a constant which depends, ultimately, only on the density of the original set A'.)

Why does this follow from Theorem 7.2? Well, that theorem implies that a large subset of Γ is contained in a smallish multidimensional (but low-dimensional) arithmetic progression $P \subset \mathbb{Z}_N^2$. Such a progression will be a set of the form

$$\Big\{(x_0,y_0) + \sum_{i=1}^d a_i(x_i,y_i) : 0 \leq a_i < k_i \ (i = 1,2,\ldots,d)\Big\},$$

which is the graph of a linear-like function ψ that takes $x_0 + \sum_{i=1}^d a_i x_i$ to $y_0 + \sum_{i=1}^d a_i y_i$. Now the size of P is proportional to N, so at least one of the k_i will be of size $m \geq N^{1/d}$, where d is the dimension of P. Now P can be partitioned into one-dimensional progressions obtained by fixing all the (x_j,y_j) apart from (x_i,y_i),

and these all have size m. These progressions are of the form $\{(u,v)+a_i(x_i,y_i): 0\leq a_i<m\}$, which, provided x_i is non-zero (which can easily be shown), is the graph of a linear function defined on a one-dimensional arithmetic progression in $\mathbb{Z}_N$ of length m. By averaging, at least one of these small graphs has a substantial intersection with Γ, which is the same as saying that ϕ has a small part which is linear.

8. Progressions of Length Four.

We have now shown that if $A\cap(A+k)^\sim(\phi(k))$ is large for many values of k then ϕ resembles a linear function. To simplify the exposition, let us now assume that ϕ *is* linear, and see what we can deduce. The constant term is unimportant, so let us suppose, for convenience, that $\phi(k)=2ck$ for every k, for some constant $c\in\mathbb{Z}_N$. Recall from the beginning of the proof of Proposition 6.1 that this implies (indeed, is equivalent to) the inequality

$$\sum_k\sum_{s,u}f(s)\overline{f(s-k)f(s-u)}f(s-k-u)\omega^{-\phi(k)u}\geq\alpha N^3\ ,$$

which in our case becomes

$$\sum_k\sum_{s,u}f(s)f(s-k)f(s-u)f(s-k-u)\omega^{-2cku}\geq\alpha^3N^3\ .$$

Using the identity

$$2ku=s^2-(s-k)^2-(s-u)^2+(s-k-u)^2$$

we can deduce that

$$\sum_r\sum_{a,b,c,d}f(a)f(b)f(c)f(d)\omega^{-r(a-b-c+d)}\omega^{-c(a^2-b^2-c^2+d^2)}\geq\alpha^3N^4\ ,$$

or in other words that

$$\sum_r\Big|\sum_s f(s)\omega^{-cs^2}\omega^{-rs}\Big|^4\geq\alpha^3N^4\ .$$

Now $\sum_s f(s)\omega^{-cs^2}\omega^{-rs}$ is just $\hat{g}(r)$, where h is the function $s\mapsto f(s)\omega^{-cs^2}$. By the implication of (iii) from (iv) in Lemma 2.2, we therefore deduce that for some value of r we have the lower bound

$$\Big|\sum_s f(s)\omega^{-cs^2}\omega^{-rs}\Big|\geq\alpha^{3/2}N$$

or in other words that A exhibits quadratic bias of a particularly strong kind.

If ϕ is not linear but does at least have a small linear part, such as is given by Corollary 7.10, then a similar result can be proved, but the conclusion is weaker. The proof is similar as well, but messier. Since the main idea is contained in the above argument, I shall simply state the result that comes out.

Proposition 8.1. *Let $A\subset\mathbb{Z}_N$ have balanced function f. Let P be an arithmetic progression (in $\mathbb{Z}_N$) of cardinality T. Suppose that there exist λ and μ such*

that $\sum_{k\in P}|\Delta(f;k)\tilde{}(\lambda k+\mu)|^2 \geq \beta N^2 T$. *Then there exist quadratic polynomials* $\psi_0,\psi_1,\dots,\psi_{N-1}$ *such that*

$$\sum_s \Big|\sum_{z\in P+s} f(z)\omega^{-\psi_s(z)}\Big| \geq \beta NT/\sqrt{2}\ .$$

Since our aim is to iterate by passing to subprogressions, the fact that f exhibits only this weak form of quadratic bias is not particularly damaging.

We are now ready to give a proof, with a couple of details sketched, of Szemerédi's theorem for progressions of length four. Once again, all numbers have been put in explicitly - I hope the reader can see through them to the main idea of the argument.

Theorem 8.2. *There is an absolute constant* C *with the following property. Let* A *be a subset of* $\mathbb{Z}_N$ *with cardinality* δN. *If* $N \geq \exp\exp\big((1/\delta)^C\big)$, *then* A *contains an arithmetic progression of length four.*

Proof. Our assumption certainly implies that $N \geq 32k^2\delta^{-k}$. Suppose now that the result is false. Then Corollary 3.6 implies that A is not α-quadratically uniform, where $\alpha = (\delta/2)^{64}$. By Lemma 3.1 (in particular the implication of (i) from (v)) there is a set $B \subset \mathbb{Z}_N$ of cardinality at least $\alpha N/2$ together with a function $\phi : B \to \mathbb{Z}_N$, such that $|\Delta(f;k)\tilde{}(\phi(k))| \geq \alpha N/2$ for every $k \in B$. In particular,

$$\sum_{k\in B}|\Delta(f;k)\tilde{}(\phi(k))|^2 \geq (\alpha/2)^3N^3\ .$$

Hence, by Proposition 6.1, B contains at least $(\alpha/2)^{12}N^3$ ϕ-additive quadruples.

By Corollary 7.10, we can find a mod-N arithmetic progression P of size at least $N^{2^{-32000}\alpha^{30000}}$ and constants $\lambda,\mu \in \mathbb{Z}_N$ such that

$$\sum_{k\in P}|\Delta(f;k)\tilde{}(\lambda k+\mu)|^2 \geq 2^{-16000}\alpha^{15000}|P|N^2\ .$$

Therefore, by Proposition 8.1, we have quadratic polynomials $\psi_0,\psi_1,\dots,\psi_{N-1}$ such that

$$\sum_s \Big|\sum_{z\in P+s} f(z)\omega^{-\psi_s(z)}\Big| \geq \beta N|P|/\sqrt{2}$$

where $\beta = 2^{-16000}\alpha^{15000}$.

By a simple averaging argument we can find a partition of $\mathbb{Z}_N$ into mod-N arithmetic progressions $P_1,\dots,P_M$ of length $|P|$ or $|P|+1$ and also a sequence $\psi_1,\dots,\psi_M$ (after renaming) of quadratic polynomials such that

$$\sum_{j=1}^M \Big|\sum_{z\in P_j} f(z)\omega^{-\psi_j(z)}\Big| \geq \beta N/2\ .$$

(Each P_j is either a translate of P or a translate of P extended by one point. Because of the small extensions we have changed $\sqrt{2}$ to 2.)

Using Lemma 2.3, one can partition each P_j into genuine arithmetic progressions (rather than just mod-N ones), obtaining a refinement $Q_1,\dots,Q_L$, which

automatically satisfies an inequality of the form

$$\sum_{j=1}^{M}\Big|\sum_{z\in Q_j} f(z)\omega^{-\psi_j(z)}\Big| \geq \beta N/2\,.$$

Once again, we have renamed the functions ψ_j. It turns out that one can take $L \leq N^{1-2^{-32002}\alpha^{30000}}$. Next, the results of Section 5 give us a further refinement of $Q_1,\dots,Q_L$ into arithmetic progressions $R_1,\dots,R_H$ such that the functions $\psi(j)$ are approximately constant on each R_h, and can therefore be ignored. More precisely, one obtains the inequality

$$\sum_{i=1}^{H}\Big|\sum_{s\in R_i} f(s)\Big| \geq \beta N/4$$

where H is at most $N^{1-2^{-32010}\alpha^{30000}}$. Finally, an averaging argument gives us an arithmetic progression R of cardinality at least $\beta N^{2^{-32010}\alpha^{30000}}$ such that $\sum_{s\in R} f(s) \geq \beta|R|/16$. This implies that the cardinality of $A\cap R$ is at least $|R|(\delta+2^{-16004}\alpha^{15000})$. Recalling that $\alpha = (\delta/2)^{64}$, we find that the density of A has gone up from δ in $\mathbb{Z}_N$ to at least $\delta(1+(\delta/2)^{980000})$ inside the arithmetic progression R.

We now iterate this argument as in the proof of Roth's theorem. The iteration can be performed at most $(\delta/2)^{-1000000}$ times, and at each step the value of N is raised to a power which exceeds $(\delta/2)^{2000000}$. It is not hard to check that N will always remain sufficiently large for the argument to work, as long as the initial value of N is at least $\exp\exp(\delta^{-C})$, where C can be taken to be 2000000. □

An alternative formulation of the condition on N and δ is that δ should be at least $(\log\log N)^{-c}$ for some absolute constant $c > 0$. We have the following immediate corollary.

Corollary 8.3. *There is an absolute constant $c > 0$ with the following property. If the set $\{1,2,\dots,N\}$ is coloured with at most $(\log\log N)^c$ colours, then there is a monochromatic arithmetic progression of length four.* □

9. Progressions of length greater than four.

As I said in the introduction, dealing with arithmetic progressions of length five or more is considerably more difficult than dealing with progressions of length four, even though similar ideas are used. In this section I shall outline the main

steps of the proof for progressions of length five, and in the next one I shall say more about the step that causes by far the most extra difficulty.

An outline of the proof.

A. By Corollary 3.6, if a set A fails to contain an arithmetic progression of length five, then it fails to be α-cubically uniform for some α that depends only on the density of A.

B. By the definition of α-cubically uniform (or at least, by one of the equivalent definitions), this means that there exists a set $B \subset \mathbb{Z}_N^2$ of size at least αN^2 and a function $\phi : B \to \mathbb{Z}_N^2$ such that $\Delta(f;k,l)\hat{}(\phi(k,l)) \geq \alpha N$ for every $(k,l) \in B$. Here, f is the balanced function of A (which is defined shortly after Lemma 2.1).

C. Just as the function ϕ that occurred in the proof for progressions of length four turned out to have a small linear piece, so this function ϕ has a small *bilinear* piece. To be precise, there exist arithmetic progressions P and Q of size $N^{c(\alpha)}$ with the same common difference, and a bilinear function $\beta : P \times Q \to \mathbb{Z}_N$, such that $\phi(x,y) = \beta(x,y)$ for at least $c'(\alpha)|P||Q|$ values of $(x,y) \in P \times Q$.

D. To simplify the exposition, let us now suppose that $P = Q = \mathbb{Z}_N$ and β is the function $(x,y) \mapsto 6cxy$ for some constant $c \in \mathbb{Z}_N$. Since $\Delta(f;k,l)\hat{}(\phi(k,l)) \geq \alpha N$ for every $(k,l) \in \mathbb{Z}_N^2$, we have the inequality

$$\sum_{k,l} |\Delta(f;k,l)\hat{}(6ckl)|^2 \geq \alpha N^4$$

Expanding out the modulus squared as

$$\sum_{s,t} \Delta(f;k,l)(s)\overline{\Delta(f;k,l)(t)}\omega^{-6ckl(s-t)}$$

and then making the substitution $m = s - t$, we find that

$$\sum_s \sum_{k,l,m} \Delta(f;k,l,m)(s)\omega^{-6cklm} \geq \alpha N^4 \ .$$

If we now use the identity

$$6klm = \sum_{\epsilon_1,\epsilon_2,\epsilon_3} (s - \epsilon_1 k - \epsilon_2 l - \epsilon_3 m)^3 \ ,$$

where the sum is over the eight triples $(\epsilon_1,\epsilon_2,\epsilon_3)$ with $\epsilon_i = 0$ or 1, then, writing C for the operation of complex conjugation, we can deduce that

$$\sum_s \sum_{k,l,m} \prod_{\epsilon_1,\epsilon_2,\epsilon_3} C^{\epsilon_1+\epsilon_2+\epsilon_3}\Big(f(s - \epsilon_1 k - \epsilon_2 l - \epsilon_3 m)\omega^{-c(s-\epsilon_1 k-\epsilon_2 l-\epsilon_3 m)^3}\Big) \geq \alpha N^4 \ .$$

This does not immediately tell us that f is cubically biased, as would be the most obvious generalization of the proof for progressions of length four. However, what it does tell us is that the function $x \mapsto f(x)\omega^{-cx^3}$ fails to be *quadratically* uniform. Indeed, it is precisely the negation of equivalence (ii) of Lemma 3.1 for this function.

E. If P and Q are not equal to $\mathbb{Z}_N$, then a similar statement can be proved, but this time it involves partitioning $\mathbb{Z}_N$ into subprogressions P_i and finding cubic functions κ_i such that $x \mapsto f(x)\omega^{-\kappa_i(x)}$ is not quadratically uniform inside P_i (in a sense that can, with a little effort, be made precise).

F. We can now apply the results proved earlier about quadratic non-uniformity (in particular, Proposition 6.1, Corollary 7.5 and Proposition 8.1) to find a partition of $\mathbb{Z}_N$ into subprogressions P_{ij} for each of which there is a quadratic function q_{ij} such that $\sum_{x\in P_{ij}} f(x)\omega^{-\kappa_i(x)-q_{ij}(x)}$ is, on average, of magnitude comparable to $|P_{ij}|$.

G. As in the proof of Theorem 8.2, one can then use the results of Section 5 to pass to further subprogressions in which the cubic functions $\kappa_i + q_{ij}$ are approximately constant, and then ignore these functions. Finally, by an averaging argument one finds one of these subprogressions, R say, with the property that $\sum_{x\in R} f(x) \geq c''(\alpha)|R|$, which implies that $|A \cap R| \geq (\delta + c''(\alpha))$.

H. It turns out that $c''(\alpha)$ is a power of α (albeit a very large one) and therefore that the above argument can be iterated to prove that every subset of $\{1, 2, \dots, N\}$ of size at least $N/(\log\log N)^\gamma$ contains an arithmetic progression of length at least five. Here, γ is a large power of α.

10. Finding a small bilinear piece.

Of the steps outlined in the previous section, most are straightforward generalizations of the corresponding steps for progressions of length four. Step E is annoyingly messy, though not fundamentally difficult. The genuinely troublesome step is C. Why should this be so hard? The answer is that we have left behind the comfortable world of linearity, and in fact the proof involves quadratic functions in a fundamental way.

One might try to reason as follows. If one considers the function $\Delta(f;k,l)$ for fixed k, then it is simply the function $\Delta(g_k;l)$, where g_k is the function $D(f;k)$. If $\Delta(f;k,l)(\phi(k,l))$ is often large, then for many values of k we must find that $\Delta(g_k;l)(\phi(k,l))$ is often large. But then, by the results of Sections 6 and 7, it follows that the function $l \mapsto \phi(k,l)$ is linear-like. In other words, fixing k often gives a linear-like function in l and, by symmetry, fixing l often gives a linear-like function in k. This suggests that ϕ is bilinear-like.

Unfortunately, it is difficult to get the linear behaviours of ϕ in the individual variables to interact with one another. The following example illustrates one sort of problem that can occur. Let λ be an arbitrary function from $\mathbb{Z}_N$ to $\mathbb{Z}_N$, and define

$$\phi(x,y) = \begin{cases} \lambda(x)y & 0\leq x\leq y<N \\ x\lambda(y) & 0\leq y<x<N\,. \end{cases}$$

There are certainly many additive quadruples in each variable, and plenty of resulting linearity, but if λ does not have special additivity properties, then the quadruples with x fixed do not mix with those with y fixed and there is nothing more to say about ϕ, and in particular no restriction of ϕ that looks bilinear.

This example can be dealt with by the observation that not only ϕ but any restriction of ϕ to a reasonably large set should be linear-like in each variable. However, there is another phenomenon that causes far more difficulty. Let us informally call a function *quasilinear* if it resembles a low-dimensional linear function (such as, for example, the function defined at the end of §6). A serious complication arises even if we know for every x that $\phi(x,y)$ is quasilinear in y for every x and vice-versa.

It is tempting to suppose that one might be able to find a large subset $B' \subset B$, and numbers $x_0, x_1, \dots, x_d$, $r_1, \dots, r_d$, $y_0, y_1, \dots, y_d$, $s_1, \dots, s_d$ and $(c_{ij})_{i,j=0}^d$ such that the restriction of ϕ to B' was of the form

$$\phi\Big(x_0 + \sum_{i=1}^d a_i x_i, y_0 + \sum_{i=1}^d b_j y_j\Big) = \sum_{i,j=0}^d c_{ij} a_i b_j$$

for $0 \leq x_i < r_i$ and $0 \leq y_j < s_j$.

However, this would imply that one could find a small "common basis" for all the functions $y \mapsto \phi(x,y)$ (and similarly the other way round) and a simple example shows that such a statement is too strong. Indeed, let ψ be a non-trivial (i.e., non-linear) quasilinear function from $\mathbb{Z}_N$ to $\mathbb{Z}_N$. (For definiteness one could let $\psi(z) = z \pmod m$ for some m near $\sqrt{N}$.) Define $\phi(x,y)$ to be $\psi(xy)$. The natural bases for the functions $y \mapsto \psi(xy)$ are all completely different, and there is no small basis that can be used for all (or even a large proportion) of them. We shall not prove this here, but it rules out simple proofs, or even definitions, of the bilinearity of ϕ.

It is not easy to say much about how we do actually go about the task. Here is a very vague sketch.

(1) Given any $h \in \mathbb{Z}_N$, define $\phi_h(x,y)$ to be $\phi(x,y+h) - \phi(x,y)$ (with the obvious convention that this is undefined unless both $\phi(x,y+h)$ and $\phi(x,y)$ are defined).

(2) Prove a lemma, along similar lines to Proposition 6.1, to the effect that many of the ϕ_h have many additive quadruples. Notice that since the ϕ_h are defined in terms of more than one row of $\mathbb{Z}_N$, knowledge about the ϕ_h collectively has the effect of 'linking' these rows.

(3) Use another lemma to pass to a subset B' of B with the property that, for almost every ϕ_h, almost every additive quadruple where ϕ_h is defined (now defining ϕ_h only at those (x,y) such that both (x,y) and $(x,y+h)$ belong to B') is ϕ_h-additive. That is, if $x_1 + x_2 = x_3 + x_4$ then we tend to expect

$$\phi_h(x_1,y) + \phi_h(x_2,y) = \phi_h(x_3,y) + \phi_h(x_4,y)$$

(4) Now define functions g_h by setting $g_h(x)$ to be the number of y such that both (x,y) and $(x,y+h)$ belong to B'. Another proposition similar to Proposition 6.1 can be used to show that the Fourier coefficients of the functions g_h are related. Specifically, if σ is any function such that $\hat{g}_h(\sigma(h))$ is large for many values of h, then there are many σ-additive quadruples, and hence, by the results of Section 7, σ is linear-like.

(5) The functions ϕ_h are also linear-like. Moreover, the Fourier coefficients of the functions g_h are closely related to the multidimensional arithmetic progressions that serve as the 'domains' of the ϕ_h.

(6) Using the results of Section 5, the above facts and many averaging arguments, one can find arithmetic progressions P and Q with the same common difference (this detail is important and turns out to require quadratic methods) such that, for many values of $h \in P$, ϕ_h behaves linearly in y for many $y \in Q$.

(7) By further averaging arguments, we can find P' and Q' such that for many $x \in P'$, $\phi(x,y)$ behaves linearly for many $y \in Q'$.

(8) Because ϕ is linear-like in x as well as in y, we can deduce that the 'gradients' of the functions $x \mapsto \phi(x,y)$ are often related in a linear way. This gives us the required bilinear piece of ϕ.

I do not expect the above sketch to be fully comprehensible, but I hope that it gives some flavour of the argument, and some idea of its complexity.

11. Some unsolved problems in additive/combinatorial number theory.

There are many interesting open problems that it may now be possible to tackle, though they all seem to be challenging. In the remainder of this paper, we shall discuss some of these, and in this way give a brief survey of the more general area of number theory of which Szemerédi's theorem forms a part. There is no completely satisfactory name for this area: it lies at the interface between additive number theory, harmonic analysis and combinatorics. Perhaps one could characterize it negatively as that corner of number theory where neither algebraic methods nor the Riemann zeta function and its generalizations play a central role. Alternatively, one could describe it as the part of number theory that immediately appeals to combinatorialists, even if they cannot rely exclusively on combinatorics to solve its problems.

Problem 11.1.

It is known by Furstenberg's methods that the following multidimensional version of Szemerédi's theorem holds: for every δ, k and d and any finite subset $K \subset \mathbb{Z}^d$ there exists N such that every subset $A \subset \{1,2,\dots,N\}^d$ of size at least δN^d contains a homothetic copy $a + bK$ of K. However, there is no proof known that gives any bound for N. In fact, even when $d = 2$ and A is the set $\{(0,0),(0,1),(1,0)\}$ there is no good bound known.

In fact, the best bound so far was discovered very recently by Solymosi [So]. His proof relies on a curious lemma of Ruzsa and Szemerédi. In order to explain it, we must first recall some terminology from graph theory. Let G be a bipartite graph whose edges join the two (disjoint) vertex sets X and Y. A *matching* in G is defined to be a set of edges (x_i,y_i) with $x_i \in X$, $y_i \in Y$ and all the x_i and y_i distinct. In other words, thinking of each edge as the set consisting of its two end-vertices, all edges in a matching are required to be disjoint. Given any graph G, an *induced subgraph* of G is any graph H whose vertex set W is a subset of the vertex set V of G and whose edges consist of all pairs $\{x,y\} \subset W$ such that x and y are joined in G. (This is very different from the more general notion of a subgraph, which is simply any graph formed by a subset of the edges of G.)

Returning to the bipartite case, an *induced matching* is, of course, a matching that happens also to be an induced subgraph of G. To find an induced matching, one must choose subsets $\{x_1,\dots,x_k\} \subset X$ and $\{y_1,\dots,y_k\} \subset Y$ such that x_i is joined to y_j in G if *and only if* $i = j$. If G has many edges, then one naturally

expects induced matchings to be hard to come by since they have very few edges. The Ruzsa-Szemerédi lemma provides some confirmation of this.

Lemma 11.2. *Let C be a constant and let G be a bipartite graph with vertex sets X and Y of size n. Suppose that the edges of G can be expressed as a union of Cn induced matchings. Then G has $o(n^2)$ edges.*

What is curious about this lemma is that its conclusion is so weak. If G had cn^2 edges for some constant $c > 0$, and could be written as a union of only Cn induced matchings, then the average size of a matching would be cn/C. The number of vertices of a typical matching would therefore be within a constant of maximal, and there would be almost no edges between these vertices (because the matching is induced). Thus, G would be full of enormous holes wherever its edges concentrated.

One might expect such thoughts to lead to a fairly easy proof that the number of edges in G was at most n^α for some $\alpha < 2$. However, not only do they not do so, but this conclusion is not even true, as it would imply an upper bound for Roth's theorem which is better than the best known lower bound. See the discussion of problem 11.6 below for more details on this point.

The proof of Lemma 11.2 is not too hard, but it relies on Szemerédi's famous regularity lemma, which is an extremely useful graph-theoretic tool but which gives rise to bounds of tower type. This is the reason that the Ruzsa-Szemerédi lemma is usually not stated in a more quantitative form. What their proof actually shows is that the number of edges can be at most $n^2/f(n)$ where f is a function that grows roughly as slowly as $\log^*(n)$. (This is defined as the number of times you must take logarithms in order to get n down to 1.)

Now let A be a dense subset of $\{1, 2, \dots, N\}^2$. Armed Solymosi defines a bipartite graph G with vertex sets X and Y both copies of $\{1, 2, \dots, N\}$ and joins x to y if and only if $(x, y) \in A$. In addition, for each $d \in \{-(N-1), -(N-2) \dots, N-1\}$ he defines a matching M_d to consist of all edges (x, y) with $x - y = d$. Since G has more than $o(n^2)$ edges and these have been written as a union of fewer than $2N$ matchings, the Ruzsa-Szemerédi lemma implies that not all these matchings are induced. This, it can be checked, implies that A contains a configuration of the form $\{(x, y), (x + d, y), (x, y - d)\}$. If we therefore apply the argument to a suitable reflection of A, we obtain a triangle of the form $\{(x, y), (x + d, y), (x, y - d)\}$ as required by the theorem. (Irritatingly, there seems to be no obvious way to force d to be positive, so the rightangle may be the bottom left corner or it may be the top right.)

Problem 11.3

Although a power-type bound for the Ruzsa-Szemerédi lemma does not hold, it is extremely unlikely that a function like $n^2/\log^*(n)$ gives the correct order of magnitude. Given that the lemma has such interesting direct consequences, there is strong motivation for the following problem: find a proof of the Ruzsa-Szemerédi lemma which does not use Szemerédi's regularity lemma, and which (consequently, one imagines) gives a significantly better bound.

Problem 11.4

Recently, Bergelson and Leibman proved the following beautiful 'polynomial Szemerédi theorem' [BL]. For any $\delta > 0$ and any collection $p_1, \dots, p_k$ of polynomials that have integer coefficients and vanish at zero, there exists N such that every set $A \subset \{1, 2, \dots, N\}$ of size at least δN contains a subset of the form $\{a + p_1(d), a + p_2(d), \dots, a + p_k(d)\}$. Letting p_i be the polynomial id, one immediately recovers Szemerédi's theorem. Once again, this theorem is known only by the ergodic theory method and hence no bound for N is known.

Even guaranteeing the existence of a subset of the form $\{a, a+d^2\}$ is not trivial, but for simple examples like this there are explicit bounds, due to Sárközy [S] and others [PSS]. (In its qualitative form, this result was discovered independently of Sárközy by Furstenberg.) The analytic proof of Szemerédi's theorem outlined in this paper suggests that it ought to be possible to prove a quantitative version of the Bergelson-Leibman theorem as well. This, one might hope, could be developed from the proof of a simple case such as $\{a, a+d^2\}$ rather as the proof of Szemerédi's theorem has its roots in Roth's much simpler argument for progressions of length three.

One difficulty with this project is that Sárközy's argument is *not* all that simple. Very recently, however, Green [Gre1] has discovered a proof of the Furstenberg-Sárközy theorem which, though giving a worse bound than Sárközy obtains, has the merit of being simpler and, more importantly, closely analogous to the proof of Roth's theorem. If a quantitative version of the Bergelson-Leibman theorem is ever discovered, it will probably begin with Green's argument.

Not too surprisingly, progress has so far been modest. Green's paper contains a highly ingenious quantitative proof that if $\delta > 0$, N is sufficiently large and A is a subset of $\{1, 2, \dots, N\}$ of size at least δN, then A must contain an arithmetic progression of length three whose common difference is a sum of two squares. This restriction on the common difference forces him to use quadratic methods similar to those needed in this paper to deal with progressions of length four.

Just as Szemerédi's theorem can be thought of as the density version of van der Waerden's theorem, so the Bergelson-Leibman theorem is the density version of the following colouring statement.

Theorem 11.5. *Let $p_1, \dots, p_k$ be polynomials that vanish at zero and have integer coefficients. Then for every positive integer r there exists N (depending only on r and $p_1, \dots, p_k$) such that, however the set $\{1, 2, \dots, N\}$ is coloured with r colours there exist a and $d \neq 0$ such that all the numbers $a + p_i(d)$ have the same colour.*

Bergelson and Leibman proved this theorem first and then applied Furstenberg's methods to obtain the stronger density statement. Even their proof of the colouring statement used ergodic theory, so it was of considerable interest when Walters [W] found a purely combinatorial proof of Theorem 11.5 which was very much in the spirit of van der Waerden's original arguments. It seems not to be possible to find a 'Shelah-ization' of Walters's proof, so it is still an open problem whether the bounds for Theorem 11.5 can be made primitive recursive.

Problem 11.6

The following famous question was discussed in the introduction: do the primes contain arbitrarily long arithmetic progressions? There are two obvious approaches to it. The first, and more ambitious, is to improve the density bound in Szemerédi's

theorem enough to show that a density of $(\log n)^{-1}$ is sufficient. (In fact, it is an amusing exercise to show that even a density of $C \log\log n/\log n$, for a certain absolute constant C, will do.) The second approach, which at the moment seems more realistic, is to start with Vinogradov's methods, which can be used to prove that the primes contain infinitely many progressions of length three, and try to generalize them in the way that the methods of this paper generalize the proof of Roth's theorem.

There are at least two major obstacles to carrying out this very natural programme. The first, which seems to be more fundamental, is that any use of Freiman's theorem will be for a constant C which is comparable to $\log n$. Since the best known estimate for the dimension of the arithmetic progression given by the theorem is itself comparable to C (see Problem 11.9 for further discussion of this), and since not much can be said about a $(\log n)$-dimensional arithmetic progression, it appears that progress with the primes will have to wait until there has been progress in our understanding of Freiman's theorem.

The second obstacle is related to the way we *used* Freiman's theorem. When proving Szemerédi's theorem, one can afford to pass to a small subprogression and start again, as long as the set is reasonably dense. However, if one wishes to use the structure of the set of primes, then this move is ruled out: next to nothing is known about the structure of the primes when they are restricted to an arithmetic progression of length, say, $N^{1/100}$. One can imagine ways round this difficulty: it ought to be possible to strengthen the argument of Section 8 to deduce from the quadratic non-uniformity of a function f not just that $\sum_{x\in P} f(x)\omega^{q(x)}$ is unexpectedly large for some quadratic function q and some smallish arithmetic progression P, but that $\sum_x f(x)\omega^{\psi(x)}$ is large, where now the sum is over the whole of $\mathbb{Z}_N$ and ψ is some sort of 'multidimensional quadratic form'. A more global statement such as this should be easier to disprove in the case of the primes using standard methods. Unfortunately, it is not easy even to formulate an appropriate statement, let alone prove it.

Problem 11.7

The following is probably the most famous of all the unsolved problems of Erdős. Let X be a subset of $\mathbb{N}$ with the property that $\sum_{x\in X} x^{-1} = \infty$. Does X necessarily contain arithmetic progressions of every length? It is not known even whether X must contain an arithmetic progression of length three. If the problem has a positive answer, then it implies that the primes contain arbitrarily long progressions.

Although the form of the conjecture is amusingly neat, one should not be misled into thinking that there is anything particularly natural about the sum of reciprocals. It is an easy exercise to show that if $\sum_{x\in X} x^{-1} = \infty$ then for any $\epsilon > 0$ the size of $X \cap \{1, 2, \dots, N\}$ is at least $N/(\log N)^{1+\epsilon}$ infinitely often. Thus, Erdős's conjecture would follow if one could show that a density of $1/(\log N)^{1+\epsilon}$ was enough in Szemerédi's theorem. Conversely, if a sequence of sets $A_1, A_2, \dots$ can be found, where each A_m is a subset of $\{1, 2, 3, \dots, 2^m\}$ of size at least $2^m/m$ not containing an arithmetic progression of length k, then the union $X = (A_1 + 2) \cup (A_3 + 2^3) \cup (A_5 + 2^5) \cup \dots$ still contains no arithmetic progression of length k even though $\sum_{n\in X} n^{-1} = \infty$. Thus, Erdős's conjecture follows from, and is roughly equivalent to, the statement that Szemerédi's theorem is true with a density significantly better

than $1/\log n$. I say 'roughly equivalent' because it is conceivable that, although a density of $1/\log n$ is insufficient for Szemerédi's theorem, the counterexamples are so few and far between that it is not possible to put them together to obtain a set X such that $\sum_{n\in X} n^{-1} = \infty$. For example, this would be true if a density of $\log n$ was *usually* sufficient, but not quite always, owing to a strange construction that worked only inside intervals of length of the form 2^{m^2}. Of course, not only is such a scenario highly unlikely, it would also not matter in the case of sets such as the primes, which have density $1/\log n$ not just sporadically, but all the time.

Problem 11.8

These observations show that the prettiness of Erdős's conjecture is somewhat artificial, and that the real question is the more prosaic (but still fascinating) one about the correct density in Szemerédi's theorem. Given that progressions of length three are much easier to handle than longer ones, it is very frustrating that the following special case of the problem is still wide open: what is the correct bound for Roth's theorem?

In Section 2 we saw that a density of $C/\log\log N$ is enough to guarantee an arithmetic progression of length three. This bound was improved by Szemerédi [Sz3] and Heath-Brown [H-B] to $(\log N)^{-c}$ for an absolute constant $c > 0$. The best known result in this direction was obtained recently by Bourgain [Bou], who showed that a density of $C\log\log N/(\log N)^{1/2}$ was enough. The reason Bourgain obtains a much stronger bound than Roth is, very roughly, as follows. The main source of inefficiency in Roth's argument is the fact that one passes many times to a subprogression of size the square root of what one had before. This means that the iteration argument is very costly. Moreover, at each stage of the iteration, one obtains increased density on a mod-N arithmetic progression of *linear* size and simply discards almost all of this information in the process of restricting to a 'genuine' arithmetic progression.

Bourgain does not throw away information in this way. Instead, he tries to find increased density not on arithmetic progressions but on translates of *Bohr neighbourhoods*, which are sets of the form $\{x \in \mathbb{Z}_N : r_i x \in [-\delta_i N, \delta_i N]\}$. Note that these sets are just intersections of a few mod-N arithmetic progressions. Roughly speaking, if a set A is not evenly distributed inside a Bohr neighbourhood B, then, using a large Fourier coefficient of $A \cap B$, one can pick out a new mod-N arithmetic progression P such that the density of A inside $B \cap P$, which is still a Bohr neighbourhood, is larger. The reason this approach can be expected to work is that Bohr neighbourhoods have a great deal of arithmetic structure: indeed, they are rather similar to multidimensional arithmetic progressions. I should make clear that this sketch of Bourgain's method, although it conveys the basic idea, is not quite an accurate portrayal of what he actually does. The technicalities involved in getting something like this idea to work are formidable and Bourgain's paper is a tour de force.

As I have said, the discrepancy between this bound and the best known lower bound is very large. The lower bound comes from a construction of Behrend [Be]. It was published in 1946, and nobody has found even the smallest improvement. Since the construction is beautiful, simple and gives an important insight into why the problem is difficult, it is worth giving in full.

To begin with, let us construct a different object. Let m and d be parameters to be chosen later, and let us search for a subset A of the grid $\{0,1,2,\dots,m-1\}^d$ containing no arithmetic progression of length three. (This means a set of three points x, y and z in the grid such that $x+z=2y$.) A simple way to do this is to choose a positive integer t and let A_t be the set of all x such that $x_1^2+\dots+x_d^2=t$. Since all points in A_t then lie on the surface of a sphere, it is clear that A_t contains no arithmetic progression (or even a set of three collinear points). Furthermore, since A_t is only ever non-empty for $d \le t \le m^2 d$, and every point in the grid lies in some A_t, averaging tells us that there exists a t such that A_t has cardinality at least $m^d/m^2 d$.

The next observation is that the grid can be embedded into $\mathbb{N}$ in such a way that arithmetic progressions of length three are preserved and no new ones are created. More precisely, given a point x in $\{0,1,2,\dots,m-1\}^d$, let $\phi(x)$ be the positive integer obtained by thinking of x as a number written backwards in base $2m$. (In other words, $\phi(x)=\sum_{i=1}^d x_i(2m)^{i-1}$.) Then it is not hard to check that $\phi(x)+\phi(z)=2\phi(y)$ if and only if $x+z=2y$. (It is to obtain the 'only if' that we use base $2m$ rather than the more obvious base m.)

Furthermore, the range of ϕ is contained in an interval of length $(2m)^d$. Therefore, we can use the map ϕ to take the set A_t to a subset A of such an interval, where A has size at least $m^d/m^2 d$ and contains no arithmetic progression of length three. All that remains is to optimize the choice of m and d given that $(2m)^d=N$. It turns out that a good choice is to set $d=\sqrt{\log N}$, which results in a subset A of $\{1,2,\dots,N\}$ with no arithmetic progression of length three and with cardinality $N\exp(-c\sqrt{\log N})$.

It is perhaps easier to see how far this bound is from Bourgain's upper bound if we state the bounds in the following equivalent way. For a fixed $\delta>0$, let $D=\delta^{-1}$. Then the N needed by Bourgain to guarantee that every subset of $\{1,2,\dots,N\}$ of size at least δN contains an arithmetic progression of length three is $\exp(cD^2\log D)$, whereas Behrend's construction gives a counterexample when $N=\exp(C(\log D)^2)$.

In view of the apparent weakness of the Behrend bound, why is it regarded as so interesting? The main reason is that, as mentioned at the beginning of Section 2, it disproves a very natural conjecture (which at one time was even made by Erdős and Turán). This conjecture is that a density of $CN^{-\alpha}$ is sufficient to guarantee a progression of length k, for some $\alpha>0$ depending only on k. This is the sort of bound one would expect from the general heuristic principle that probabilistic arguments always do best. This principle is simply wrong in the case of Szemerédi's theorem.

This fact is interesting in itself, but it also has interesting metamathematical consequences. If one is trying to improve the upper bound, one can immediately rule out several potential arguments on the grounds that, if they worked, they would give rise to power-type bounds. When planning an approach to the upper bound, it is very important that there should be some foreseeable 'unpleasantness', some difficulty that would give rise to a bound expressed by a less neat function. This shows, for example, that you will not be able to prove Roth's theorem using only a little formal manipulation of Fourier coefficients. (A different way to see this is to note that the Fourier expression that counts the arithmetic progressions of length three in A does not have to be non-negative if A, a characteristic function of a set, is replaced by a more general function.) Also, there seems little hope of a compression-type proof that successively modifies an AP-free set without decreasing

its size until eventually it is forced to have some extremal structure - simply because it is hard to imagine forcing a set to have the very particular and not wholly natural quadratic structure of Behrend's example.

Problem 11.9

We saw in the discussion of Problem 11.6 some of the motivation for the following question: what are the correct bounds for Freiman's theorem? In fact, this is a question of major importance, with potential applications to all sorts of different problems. In order to discuss bounds, it is helpful to summarize very briefly Ruzsa's proof of the theorem.

Ruzsa starts with a set A_0 such that $|A_0 + A_0| \leq C|A_0|$. Then, by a highly ingenious argument, he finds a subset $A_1 \subset A$ of proportional size such that A_1 is 'isomorphic' to a subset $A \subset \mathbb{Z}_N$, where N is also proportional to $|A_0|$ (that is, not too large). Rather than say precisely what 'isomorphic' means, let me give instead the main relevant consequence, which is that if $2A - 2A$ contains a d-dimensional arithmetic progression of a certain size, then so does $2A_0 - 2A_0$.

Further arguments of a combinatorial nature can be used to show that if $2A_0 - 2A_0$ contains a large and small-dimensional arithmetic progression, then A_0 is contained in one. (Of course, this also uses the assumption that $|A_0 + A_0| \leq C|A_0|$.)

As a result, Ruzsa shows that Freiman's theorem is (non-trivially) equivalent to the following statement: if A is a subset of $\mathbb{Z}_N$ of size δN, then $2A - 2A$ contains an arithmetic progression P of size at least $c(\delta)N$ and dimension at most $d(\delta)$.

It turns out that, from the point of view of applications, the quantity for which one would most like a good estimate is $d(\delta)$. A fairly straightforward argument, based on a technique of Bogolyubov [Bo], shows that d can be taken to be at most δ^{-2}. Very recently this was improved by Chang [C], who added some interesting refinements to Ruzsa's approach and obtained a bound of $\delta^{-1}\log(\delta^{-1})$. Almost certainly, however, this bound, which is the best known, is a long way from being best possible. This may be as low as $C\log(\delta^{-1})$, which would have very significant consequences. Even an estimate of $(\log(\delta^{-1}))^C$ would be extremely interesting - for example, it would be good enough to use for Problem 11.6.

Chang also found a much more efficient way than Ruzsa's of passing from the progression inside $2A_0 - 2A_0$ to the one containing A_0, so she obtains the following bounds for Freiman's theorem. If $|A + A| \leq C|A|$ then A is contained in a progression P of dimension at most $a(C\log C)^2$ and cardinality at most $\exp(aC^2(\log C)^3)$, where a is an absolute constant. A simple example shows that these bounds are almost best possible. (Just to make this statement clear: the bounds for the progression *containing* A are close to best possible, but it would be very interesting to improve the bounds for the progression *contained in* $2A - 2A$, which are far from best possible.)

The example is the following. Let m be a large integer and let A be the geometric progression $\{1, m, \dots, m^{k-1}\}$. (As will be clear, any set with no small additive relations would do just as well.) Then $|A| = k$ and $|A + A| = k(k+1)/2$, so we have $|A + A| \leq C|A|$ for a constant C proportional to k. Now the elements of A are independent in the following sense: if $a_1, \dots, a_k$ are integers such that $\sum_{i=1}^k a_i m^{i-1} = 0$, then at least one a_i has modulus at least $m/2$. Using this fact, it is easy to see that any arithmetic progression of dimension less than k containing

A must have cardinality at least m. Since m is not bounded by any function of k, it is impossible to prove a bound for the dimension that is better than linear in C.

Note that this example can be 'fattened up': simply replace A by $A+\{1,2,\dots,t\}$. With appropriate choices of t, k and m one can find similar examples for C and $|A|$ of any desired size. Note also that such examples have no bearing on the 'inner' progression. If you are looking for a low-dimensional progression inside $2B-2B$, where $B = A+\{1,2,\dots,t\}$ as above, then all you have to do is consider one of the k 'pieces' of B, which is an interval of length t, which shows that $2B-2B$ contains a one-dimensional progression of length comparable to $C^{-1}|A|$. In general, it seems that the weakness in the known arguments for Freiman's theorem is that they do not take into account the possibility that a set with small sumset may well have a subset with much better structure.

As for further potential applications of Freiman's theorem, here are a few problems that seem to be related. Others are listed in [F3] and [C].

Problem 11.10

Yet another beautiful question of Erdős is the following. Let A be a set of n integers and let $\epsilon > 0$. Is it true that either $A+A$ or $A.A$ (the set of all products ab with $a, b \in A$) has cardinality at least $n^{2-\epsilon}$?

The idea behind this problem is, of course, that if you try to make $A+A$ small, say by making A into an arithmetic progression, then $A.A$ will be almost maximal, whereas if you try to minimize $A.A$, say by making A a geometric progession, then $A+A$ will be almost maximal. In general, whatever you do to pull the sums together seems to drive the products apart, and vice-versa.

It will probably never be possible to use Freiman's theorem directly to solve this problem: to say anything about a set A with $|A+A| \leq |A|^{1+\alpha}$ for *any* fixed $\alpha > 0$ is way beyond what is possible at the moment, and it seems unlikely that there is a useful structural statement when, say, $\alpha = 0.99$. Nevertheless, it might be possible, and would be interesting, to show that if $|A+A| \leq |A|^{1.01}$ then $|A.A| \geq |A|^{2-\epsilon}$. The best known bound for the problem as stated is due to Elekes [E], who proved that one of $A+A$ and $A.A$ must have cardinality at least $|A|^{5/4}$. His proof used the Szemerédi-Trotter theorem [ST], which is a very useful tool in combinatorial geometry.

Problem 11.11

A similar question for which Freiman's theorem may eventually be useful is the so-called Erdős ring problem. It asks whether $\mathbb{R}$ contains a subring of dimension $1/2$ - that is, a subset A of Hausdorff dimension $1/2$ which is closed under addition and multiplication. Very roughly, this corresponds to asking about the structure of sets A of integers such that $|A+A| \leq |A|^{1+\epsilon}$. An interesting discrete version of the problem is the following. Does there exist a subset A of the field $\mathbb{F}_p$ of cardinality about $p^{1/2}$ such that both $A+A$ and $A.A$ have cardinality at most $p^{o(1)}|A|$? Such a set would be 'approximately closed' under addition and multiplication. Note that if one replaces $\mathbb{F}_p$ by $\mathbb{F}_{p^2}$ then the answer is trivially yes, so any proof would have somehow to distinguish between different kinds of finite fields. (A similar remark

applies to the original ring problem - it is not hard to find a subring of $\mathbb{C}$ of half the dimension of $\mathbb{C}$, so a proof in $\mathbb{R}$ would have to distinguish $\mathbb{R}$ from $\mathbb{C}$.)

Problem 11.12

Freiman has suggested that a good enough bound for his theorem would have a bearing on a famous problem in additive number theory: what is the correct order of magnitude of the Waring number $G(k)$? Recall that this is the smallest integer m such that every sufficiently large integer can be written as a sum of m k^{th} powers. The best known upper bound is $(1+o(1))k\log k$, due to Wooley [Wo], but it is conjectured that the correct bound is linear in k - or even, more ambitiously, that it is linear with constant 1. Note that k is a trivial lower bound.

Let N be very large and let K be the set of all k^{th} powers less than N. If m is significantly larger than k and if it is not possible to write every integer between, say, $kN/2$ and $kN/4$ as a sum of m elements of K, then the cardinalities of the sets K, $K+K$, $K+K+K$, and so on, eventually cease to be close to their maximum possible values of $N^{1/k}$, $N^{2/k}$, $N^{3/k}$ and so on. Indeed, at some point there must be an r such that $rK+K$ has significantly smaller cardinality than $|rK||K|$. With a very good bound for Freiman's theorem, one might possibly be able to exploit this information to obtain a contradiction - though nobody has come up with a theorem to this effect.

Problem 11.13

A *Sidon set* of integers is a set A with the property that all its sums are distinct. That is, the only solutions of the equation $x+y=z+w$ with x,y,z and w in A are the trivial ones $x+y=x+y$ or $x+y=y+x$. There are several interesting open problems connected with such sets, of a very similar flavour to the questions discussed in this paper. The most obvious three are the following.

1. How large is the largest possible Sidon subset of $\mathbb{Z}_N$?
2. How large is the largest possible Sidon subset of $\{1,2,\dots,N\}$?
3. Suppose that A is a Sidon subset of $\mathbb{N}$ such that $A\cap\{1,2,\dots,N\}$ has cardinality at least N^α for all sufficiently large N. How large can α be?

Easy counting arguments provide upper bounds for all three problems. For example, if A is a Sidon subset of $\mathbb{Z}_N$ and has cardinality m, then $A+A$ has cardinality at least $m(m+1)/2$, from which it follows that $m\leq(2N)^{1/2}$. If instead $A\subset\{1,2,\dots,N\}$ then $A+A\subset\{2,3,\dots,2N\}$ and the same argument implies that $m\leq 2N^{1/2}$. This fact, in turn, gives an upper bound of $1/2$ for α in the third question.

It is interesting to reflect on why it is that the absence of non-degenerate solutions to the equation $x+y=z+w$ gives rise, by an easy argument, to a power-type upper bound, while the absence of non-degenerate solutions to the superficially similar equation $x+z=2y$ leads to a difficult open problem for which a power-type bound is known not to hold. One reason is simply that counting argument we have just given relies on the symmetry in the first equation, and the second does not have this symmetry. A second, which is really the first reason in a different guise, becomes clear when we look at the problems on the Fourier side. Let $A\subset\mathbb{Z}_N$. As we saw in §2, the number of solutions in A to the equation $x+z=2y$ is $N^{-1}\sum_r\hat{A}(r)^2\hat{A}(-2r)$, while the number of solutions to $x+y=z+w$ is

$N^{-1}\sum_r |\hat{A}(r)|^4$. A big difference between these two expressions is that the second is automatically positive. Of course, the first is positive as well, since it counts the number of solutions to $x+z=2y$, but we know it is positive only because we know that $\hat{A}$ is the Fourier transform of a non-negative function. If f is an arbitrary real-valued function, then it is perfectly possible for $N^{-1}\sum_r \hat{f}(r)^2\hat{f}(-2r)$ to be negative, but obviously not possible for $N^{-1}\sum_r |\hat{f}(r)|^4$ to be. (See the remark towards the end of the discussion of Problem 11.8.)

Let us now consider what is known about the three questions above. Two very simple arguments show that a Sidon subset of $\mathbb{Z}_N$ can have cardinality $cN^{1/3}$. The first is simply to choose *any* maximal Sidon set $A=\{x_1,\dots,x_m\}$. If no further element can be added to A without its losing the Sidon property, then every $x\in\mathbb{Z}_N$ can be expressed in some way as $z+w-y$ with $y,z,w\in A$. Since there are at most $|A|^3$ such numbers (clearly a more careful argument will improve this estimate by a constant) it follows that $m\geq N^{1/3}$. The same argument obviously works for subsets of $\{1,2,\dots,N\}$. A similar argument shows also that $\alpha=1/3$ is possible for infinite Sidon sets: if one greedily chooses an infinite sequence $x_1,x_2,x_3,\dots$ such that each x_k is as small as possible, given that $\{x_1,\dots,x_k\}$ remains a Sidon set, then the order of magnitude of x_k is at most k^3.

Similar bounds can be obtained by the most basic form of the probabilistic argument. Suppose, for example, that A is a subset of $\mathbb{Z}_N$ with each element chosen randomly and independently with probability p. The expected size of A is pN and the expected number of non-degenerate solutions to $x+y=z+w$ is at most p^4N^3 (actually smaller by a constant because each solution is counted more than once). If $pN\geq 2p^4N^3$, then the expected value of the size of A minus the number of non-degenerate solutions to $x+y=z+w$ is at least $pN/2$. Since $p=N^{-2/3}$ satisfies this condition, we can find a set A of size at least $N^{1/3}/2$ such that, throwing away one point from each non-degenerate solution, we end up with a Sidon set of sze at least $N^{1/3}/4$. Again, with a bit more care one can improve the constant in this bound.

The best known bounds for the first two questions were obtained by Singer in 1938 using a fairly simple algebraic construction [Si]. If N happens to be of the form p^2-1 for a prime p, then he obtains a Sidon subset of $\{1,2,\dots,N\}$ of size p. By the prime number theorem, this implies a bound of $(1+o(1))N^{1/2}$ in general, which is the same order of magnitude as the trivial upper bound. In fact, even the correct constant is known, since Erdős and Turán [ET2] improved the trivial upper bound to $N^{1/2}+N^{1/4}+1$.

Despite the close agreement between these two bounds, there is considerable interest in improving them, and especially in answering the following unsolved problem: is the correct upper bound of the form $N^{1/2}+C$ for an absolute constant C? This problem, or indeed the weaker problem of finding *any* improvement of the Erdős-Turán bound, has remained open for sixty years.

The infinite problem is interestingly different from the finite one, because it is not possible, or at least not straightforwardly possible, to obtain a good lower bound by stringing together a sequence of finite examples. Indeed, until very recently the best known asymptotic size was $(N\log N)^{1/3}$, that is, only a logarithmic improvement on the trivial bound of $cN^{1/3}$. This was obtained in a seminal paper of Ajtai, Komlós and Szemerédi [AKS] - seminal partly because of its interesting

results, and partly because in it can be found the genesis of a major new technique in probabilistic combinatorics, now known as the Rödl nibble.

However, in 1998 this bound was substantially improved by Ruzsa [Ru4], who invented an astonishingly clever argument which gave a lower bound of $N^{\sqrt{2}-1+o(1)}$, the first time a power greater than $1/3$ had been obtained. Although the correct answer is almost certainly that N^α is possible for every $\alpha < 1/2$, and although such a result is unlikely to be proved by Ruzsa's method, his paper is strongly recommended for its sheer beauty and ingenuity.

In the other direction, Erdős improved the upper bound of $N^{1/2}$ by a logarithmic factor: that is, if A is an infinite Sidon set then $|A \cap \{1, 2, \dots, N\}| \leq (N/\log N)^{1/2}$ for infinitely many N. Thus, the trivial upper bound is not correct. Interestingly, a small modification to this result produces another famous open problem of Erdős.

Problem 11.14

Define a B_h-set to be a set containing no non-trivial solutions to the equation $x_1 + \cdots + x_h = y_1 + \cdots + y_h$ (so a Sidon set is a B_2-set). A simple counting argument like that for Sidon sets shows that the asymptotic size of a B_3-set cannot exceed $N^{1/3}$. However, unlike for Sidon sets, in this case it is not known whether there can be a set that achieves this trivial bound, to within a constant. There is an important technical difference between sums of two numbers and sums of three, which is that while there is a one-to-one correspondence between solutions of the equation $x + y = z + w$ and solutions of $x - y = z - w$, it is not possible to rewrite solutions to $u + v + w = x + y + z$ in terms of differences in a symmetrical way. In general, this makes problems about B_h-sets harder when h is odd.

This state of affairs can be compared with a well known problem from traditional additive number theory. While it is a classical fact that the asymptotic density of the set of sums of two squares is $c(\log N)^{-1/2}$, the problem of what happens with sums of three cubes is open. Moreover, it is conjectured that the answer is completely different, and that these numbers have positive density. (This fascinating problem does not count as 'combinatorial' in the sense in which I have been using the word, since it is clear that it will require advanced number-theoretic techniques for its solution.)

One can of course ask the corresponding finite problems for B_h-sets with $h > 2$, and for these the correct constants are no longer known. Until recently, almost all the best known upper bounds came from natural generalizations of the argument of Erdős and Turán for Sidon sets. For example, the largest B_4-subset of $\{1, 2, \dots, N\}$ was shown by Lindström [Lin] 1969 to be of size at most $(8^{1/4} + o(1))N^{1/4}$. One exception was a complicated result of Graham [Gr] which improved the naturally occurring constant for B_3-sets from $4^{1/3}$ to $(4 - \frac{1}{228})^{1/3}$. However, Green, in a paper which will appear soon [Gre2], found a genuinely new way of looking at the problem which has improved all these bounds (including Graham's), not just by

reducing the error estimates, but by actually decreasing the constant attached to the main term. In many cases these constants had stood still for over thirty years.

Bibliography.

[AKS] M. Ajtai, J. Komlós and E. Szemerédi, *A dense infinite Sidon sequence*, European J. Comb. **2** (1981), 1-11.
[BS] A. Balog and E. Szemerédi, *A Statistical Theorem of Set Addition*, Combinatorica **14** (1994), 263-268.
[Be] F. A. Behrend, *On sets of integers which contain no three in arithmetic progression*, Proc. Nat. Acad. Sci. **23** (1946), 331-332.
[BL] V. Bergelson and A. Leibman, *Polynomial extensions of van der Waerden's and Szemerédi's theorems*, J. Amer. Math. Soc. **9** (1996), 725-753.
[Bi] Y. Bilu, *Structure of sets with small sumset*, in Structure Theory of Set Addition, Astérisque **258** (1999), 77-108.
[Bo] N. N. Bogolyubov, *Sur quelques propriétés arithmétiques des presque-périodes*, Ann. Chaire Math. Phys. Kiev, **4** (1939), 185-194.
[Bou] J. Bourgain, *On triples in arithmetic progression*, Geom. Funct. Anal. **9** (1999), 968-984.
[C] M.-C. Chang, *A polynomial bound in Freiman's theorem*, submitted.
[CG] F. R. K. Chung and R. L. Graham, *Quasi-random subsets of* $\mathbb{Z}_N$, J. Combin. Theory Ser. A **61**, 64-86.
[E] G. Elekes, *On the number of sums and products*, Acta Arith. **81** (1997), 365-367.
[ET] P. Erdős and P. Turán, *On some sequences of integers*, J. London Math. Soc. **11** (1936), 261-264.
[ET2] P. Erdős and P. Turán, *On a problem of Sidon in additive number theory and on some related problems*, J. London Math. Soc. **16** (1941), 212-215.
[F1] G. R. Freiman, Foundations of a Structural Theory of Set Addition, (in Russian), Kazan Gos. Ped. Inst., Kazan (1966).
[F2] G. R. Freiman, Foundations of a Structural Theory of Set Addition, Translations of Mathematical Monographs **37**, Amer. Math. Soc., Providence, R. I., USA.
[F3] G. R. Freiman, *Structure theory of set addition*, Astérisque No. 258 (1999), 1-33.
[Fu] H. Furstenberg, *Ergodic behaviour of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204-256.
[FKO] H. Furstenberg, Y. Katznelson and D. Ornstein, *The ergodic theoretical proof of Szemerédi's theorem*, Bull. Amer. Math. Soc. **7** (1982), 527-552.
[G] W. T. Gowers, *A new proof of Szemerédi's theorem*, Geometric and Functional Analysis, to appear.
[Gr] S. W. Graham, B_h *sequences*, in Analytic Number Theory Vol. I (Allerton Park, IL, 1995), 4331-449, Progress in Mathematics **138**, Birkhäuser, Boston MA 1996.
[Gre1] B. J. Green, *On arithmetic structures in dense sets of integers*, submitted.
[Gre2] B. J. Green, *The number of squares and* $B_h[g]$ *sets*, Acta Arith., to appear.
[H-B] D. R. Heath-Brown, *Integer sets containing no arithmetic progressions*, J. London Math. Soc. (2) **35** (1987), 385-394.

[L] B. Lindström, *A remark on B_4-sequences*, Journal of Comb. Th. **7** (1969), 276-277.
[N] M. B. Nathanson, Additive Number Theory: Inverse Problems and the Geometry of Sumsets, Graduate Texts in Mathematics 165, Springer-Verlag 1996.
[PSS] J. Pintz, W. L. Steiger and E. Szemerédi, *On sets of natural numbers whose difference set contains no squares*, J. London Math. Soc. (2) **37** (1988), 219-231.
[R] K. F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 245-252.
[Ru1] I. Z. Ruzsa, *Arithmetic progressions and the number of sums*, Periodica Math. Hungar. **25** (1992), 105-111.
[Ru2] I. Z. Ruzsa, *An application of graph theory to additive number theory*, Scientia, Ser. A **3** (1989), 97-109.
[Ru3] I. Z. Ruzsa, *Generalized arithmetic progressions and sumsets*, Acta Math. Hungar. **65** (1994), 379-388.
[Ru4] I. Z. Ruzsa, *An infinite Sidon sequence*, J. Number Theory **68** (1998), 63-71.
[S] A. Sárközy, *On difference sets of sequences of integers I*, Acta Math. ACad. Sci. Hungar. **15** (1984), 205-209.
[Sh] S. Shelah, *Primitive Recursive Bounds for van der Waerden Numbers*, J. Amer. Math. Soc. **1** (1988), 683-697.
[Si] J. Singer, *A theorem in finite projective geometry and some applications to number theory*, Trans. Amer. Math. Soc. **43** (1938), 377-385.
[So] J. Solymosi, *Note on a generalization of Roth's theorem*, preprint.
[Sz1] E. Szemerédi, *On sets of integers containing no four elements in arithmetic progression*, Acta Math. Acad. Sci. Hungar. **20** (1969), 89-104.
[Sz2] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 299-345.
[Sz3] E. Szemerédi, *Integer sets containing no arithmetic progressions*, Acta Math. Hungar. **56** (1990), 155-158.
[ST] E. Szemerédi and W. T. Trotter, *Extremal problems in discrete geometry*, Combinatorica **3** (1983), 381-392.
[V] R. C. Vaughan, The Hardy-Littlewood Method (2nd ed.), Cambridge Tracts in Mathematics 125, CUP 1997.
[W] M. J. Walters, *Combinatorial proofs of the polynomial van der Waerden theorem and the polynomial Hales-Jewett theorem*, J. London Math. Soc. (2) **61** (2000), 1-12.
[We] H. Weyl, *Über die Gleichverteilung von Zahlen mod Eins*, Math. Annalen **77** (1913), 313-352.
[Wo] T. D. Wooley, *Large improvements in Waring's problem*, Ann. of Math. (2) **135** (1992), 131-164.