Some unsolved problems in additive/combinatorial number theory.

W. T. Gowers

The following article is a small modification of the last part of a longer article based on two lectures given in the Harvard/MIT series entitled ‘Current Developments in Mathematics’, which publishes proceedings every year. It takes the form of a survey of an area of number theory for which there is no completely satisfactory name. One way of describing it is to say that it lies at the interface between additive number theory, harmonic analysis and combinatorics. One could perhaps 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. Another possible definition is that it is the part of number theory that immediately appeals to combinatorialists, even if they cannot rely exclusively on combinatorics to solve its problems.

Not much background is needed to understand this survey independently of the article that preceded it, but, just to make it more self contained, let me state the theorems of Szemerédi and Freiman that will be mentioned frequently.

Szemerédi's theorem [Sz1,2,Fu,FKO,G]. *For every $\delta > 0$ and every positive integer k there exists N such that every subset A of the set $\{1, 2, \ldots, N\}$ of size at least δN contains an arithmetic progression of length k.*

The case $k = 3$ of this theorem is due to Roth [R].

Freiman's theorem [F1,2,Ru1,2,3,Bi] concerns the structure of sets with small sumsets. If A is a set of integers (or more generally a subset of an Abelian group) then the *sumset* $A+A$ is defined to be the set of all $x+y$ such that x and y belong to A. To state the theorem, we also need the notion of a multidimensional arithmetic progression. This is a straightforward generalization of the usual notion: whereas an ordinary, or one-dimensional, arithmetic progression is a set of the form $\{a + rk : 0 \leqslant r < s\}$, a d-dimensional progression is a set of the form $\{a + \sum_{i=1}^{d} r_i k_i : 0 \leqslant r_i < s_i\}$. Equivalently, it is an affine image in $\mathbb{Z}$ of an $s_1 \times \ldots \times s_d$-grid in $\mathbb{Z}^d$.

Freiman's theorem. *For every $C > 1$ there exist constants d and K such that, given any set $A \subset \mathbb{Z}$ of size N such that $A + A$ has size at most CN, there exists a progression P of dimension at most d and size at most KN such that $A \subset P$.*

It is not hard to show that if A satisfies the conclusion of Freiman's theorem, then $A+A$ has size at most 2^dKN, so in a qualitative sense the theorem characterizes sets with small sumset. However, as will be explained below, there are many situations where quantitative information is very important.

More information on Szemerédi's theorem and Freiman's theorem can be found in my (downloadable) articles on Szemerédi's theorem. Nathanson's book [N] contains a very complete account of Freiman's theorem, as well as plenty of other combinatorial/number-theoretic results.

Problem 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 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 6 below for more details on this point.

The proof of Lemma 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 $\{1, 2, \ldots, N\}^2$. Armed Solymosi defines a bipartite graph G with vertex sets X and Y both copies of $\{1, 2, \ldots, N\}$ and joins x to y if and only if $(x, y) \in A$. In addition, for each $d \in \{-(N-1), -(N-2) \ldots, N-1\}$ 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 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 4

Recently, Bergelson and Leibman proved the following beautiful 'polynomial Szemerédi theorem' [BL]. For any $\delta > 0$ and any collection $p_1, \ldots, p_k$ of polynomials that have integer coefficients and vanish at zero, there exists N such that every set $A \subset \{1, 2, \ldots, N\}$ of size at least δN contains a subset of the form $\{a+p_1(d), a+p_2(d), \ldots, a+p_k(d)\}$. 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 $\{1, 2, \ldots, N\}$ 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 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 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 5 can be made primitive recursive.

Problem 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 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 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 $X \cap \{1, 2, \ldots, N\}$ 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 $A_1, A_2, \ldots$ can be found, where each A_m is a subset of $\{1, 2, 3, \ldots, 2^m\}$ of size at least $2^m/m$ not containing an arithmetic progression of length k, then the union $X = (A_1 + 2) \cup (A_3 + 2^3) \cup (A_5 + 2^5) \cup \ldots$ 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 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 $\{0, 1, 2, \ldots, m-1\}^d$ 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 $x_1^2 + \ldots + x_d^2 = t$. 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 $d \leqslant t \leqslant m^2 d$, 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 $\{0, 1, 2, \ldots, m-1\}^d$, 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 $\{1, 2, \ldots, N\}$ 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 $\{1, 2, \ldots, N\}$ 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 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 9

We saw in the discussion of Problem 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 $|A_0 + A_0| \leqslant C|A_0|$. 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 $|A_0 + A_0| \leqslant C|A_0|$.)

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 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 $|A+A| \leqslant C|A|$ 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 $\{1, m, \ldots, m^{k-1}\}$. (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 $|A+A| \leqslant C|A|$ for a constant C proportional to k. Now the elements of A are independent in the following sense: if $a_1, \ldots, a_k$ 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 $A + \{1, 2, \ldots, t\}$. 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 $B = A + \{1, 2, \ldots, t\}$ 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 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 $|A + A| \leqslant |A|^{1+\alpha}$ 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 $|A + A| \leqslant |A|^{1.01}$ then $|A.A| \geqslant |A|^{2-\epsilon}$. 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

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 $|A + A| \leqslant |A|^{1+\epsilon}$. 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 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 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 $\{1, 2, \ldots, N\}$?

3. Suppose that A is a Sidon subset of $\mathbb{N}$ such that $A \cap \{1, 2, \ldots, N\}$ 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 $m \leqslant (2N)^{1/2}$. If instead $A \subset \{1, 2, \ldots, N\}$ then $A+A \subset \{2, 3, \ldots, 2N\}$ and the same argument implies that $m \leqslant 2N^{1/2}$. 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 $A \subset \mathbb{Z}_N$ and write $\hat{A}(r)$ for the discrete Fourier coefficient $\sum_{x \in A} \omega^{-rx}$, where $\omega = \exp(2\pi i/N)$. Then it is not hard to show (details can be found in §2 of [G]) that 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 8. Here $\hat{f}(r)$ is defined to be $\sum_x f(x)\omega^{-rx}$.)

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 $m \geqslant N^{1/3}$. 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 $pN \geqslant 2p^4N^3$, 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 $|A\cap\{1,2,\dots,N\}|\leqslant(N/\log N)^{1/2}$ 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 14

Define a B_h-set to be a set containing no non-trivial solutions to the equation $x_1+\ldots+x_h=y_1+\ldots+y_h$ (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 $\{1, 2, \ldots, N\}$ 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.

[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.

[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.

[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.

[Wo] T. D. Wooley, *Large improvements in Waring's problem*, Ann. of Math. (2) **135** (1992), 131-164.