Szemerédi's Theorem and Inverse Theory of the Gowers Norms

Talk by Freddie Manners (Notes by Sanjana Das)

March 3, 2023

§1 Introduction

Theorem 1.1 (Szemerédi)

Fix $k \geq 3$, and let $A \subseteq [N]$ with $|A| = \delta N$. If N is sufficiently large in terms of δ, then A contains a k-term arithmetic progression x, $x + h$, $x + 2h$, $\ldots$, $x + (k - 1)h$ (with $h \neq 0$).

Gowers proved the following quantitative bound on N:

Theorem 1.2 (Gowers)

If $\delta \gg (\log \log N)^{-c}$ for some constant c (or in other words, $N \gg e^{e^{\delta^{-c}}}$ for some c), then Szemerédi's theorem holds.

The case $k = 3$ was proved by Roth in 1954, with a bound similar to Gowers's. The bound in this case has been improved greatly (the current record was broken a few weeks ago).

The proof for arbitrary k was found in 1975 by Szemerédi; this proof is very difficult. In 1979 Furstenberg reproved Szemerédi's theorem using ergodic theory (this argument doesn't directly give bounds; it may be possible to extract bounds from it, but they would be quite poor). The next big milestone was in 2001, when Gowers proved the above quantitative bound. For $k \geq 5$, this is still the best bound we have (for $k = 4$ one log has been removed).

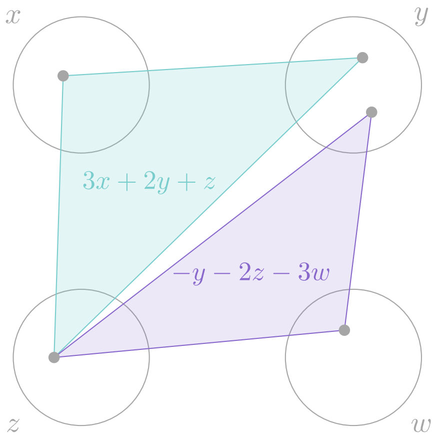
Since then, there have been many 'worse' proofs (in the sense of the quantitative bounds obtained), some of which are interesting:

- Tao used a *quantiative ergodic proof*, where by means of great cunning, you unwrap the ergodic theorem to get something combinatorial (this involves applying the regularity lemma many times).
- There is a proof by Green–Tao using the *arithmetic regularity lemma.*
- One conceptually elegant proof is by *hypergraph removal.* To illustrate this, suppose $k = 4$, so that we're looking for a 4-term arithmetic progression x, $x + h$, $x + 2h$, $x + 3h$. We can alternatively parametrize 4-term arithmetic progressions as

  $$-y - 2z - 3w, x - z - 2w, 2x + y - w, 3x + 2y + z$$

  for $x, y, z, w \in \mathbb{Z}$ (unlike the previous parametrization, this is degenerate in the sense that multiple (x, y, z, w) will produce the same progressions). (Essentially, x corresponds to arithmetic progressions whose first term is 0; y corresponds to progressions whose second term is 0; and so on.)

  Now take a 4-partite 3-uniform hypergraph, with parts corresponding to the four variables, where a triple (x, y, z) is an edge in the hypergraph precisely when $3x+2y+z$ is in A; similarly a triple (y, z, w) is an edge if and only if $-y - 2z - 3w$ is in A; and so on.



Then a 4-term arithmetic progression in A corresponds to a choice of x, y, z, and w such that all four triangles between them are edges in this hypergraph, or in other words, a clique of size 4 in the hypergraph.

So if A contradicts Szemerédi's theorem, then we have a dense hypergraph whose only cliques of size 4 correspond to degenerate progressions (one with $h = 0$). There are very few such cliques, but they are impossible to remove. The hypergraph removal lemma tells us that such hypergraphs cannot exist, which gives a contradiction.

This proof is very nice, but it gives Wowzer-type bounds. (This is a type of bound in the hierarchy of atrociously fast-growing expressions. First, there are exponentials like e^x and then iterated exponentials like e^{e^x}. Then you have tower expressions where you iterate the exponential function x times. Wowzer-type bounds are when you iterate the *tower* function x times.)

- It is also possible to deduce the theorem from density Hales–Jewett. (This also gives very bad bounds.)

§2 Gowers's Proof

§2.1 The Density Increment Strategy

The idea behind Gowers's proof is the *density increment strategy*, encapsulated by the following lemma:

Lemma 2.1
If $A \subseteq [N]$ with $|A| = \delta N$, then one of the following holds:

(a) $N \ll \delta^{-k}$.

(b) A has at least $\delta^k N^2$ arithmetic progressions of length k (up to a constant factor).

(c) There exists an arithmetic progression $P \subseteq [N]$ such that $|P| \gg N^{\delta^{O(1)}}$ and

$$\frac{|A \cap P|}{|P|} \geq \delta + \Omega(\delta^k).$$

This states that as long as N is not very small (case (a) covers very small N), either A contains loads and loads of progressions, or we can find a reasonably large patch (the progression P) where if we zoom in on this patch, the density of A increases a bit. You should think of $|P|$ as around $N^{1/2}$ or $N^{1/1000}$; the explicit exponent doesn't matter that much.

Remark 2.2. It's interesting that the patch we're zooming in on doesn't have constant density (e.g., size $\frac{1}{100}N$), and is instead much smaller. So the zooming-in argument can't give us the large number of progressions in case (b), since we're replacing N with something much smaller; but this is fine since we only need *one* progression.

First we'll see why this lemma proves Szemerédi's theorem.

Proof of Szemerédi's Theorem from Lemma. We start with $A_0 = A$ and $N_0 = N$, and perform the following recursive process. At every step, apply the lemma.

- If case (a) holds, terminate the process.
- If case (b) holds, we're done, since this case means we have a lot of arithmetic progressions. It's possible that some of the arithmetic progressions given by (b) are trivial, but we have at least $\delta^k N^2$ progressions and only δN trivial ones, and since (a) is not the case, we have $\delta^k N^2 \gg \delta N$ and there must be some nontrivial progression. (Here δ and N refer to the new parameters obtained in the recursive process.)
- If case (c) holds, then we zoom in on this progression — suppose that case (c) holds on the rth step, so we have $A_r \subseteq [N_r]$ and we've found a progression $P \subseteq [N_r]$ on which A_r has slightly greater density. Then we can drag back P to a new interval $[N_{r+1}]$ (by translating and rescaling it), and set A_{r+1} to be the points in $[N_{r+1}]$ corresponding to $A_r \cap P$ under this bijection.

We carry on the recursion until the process stops, which must happen at some point. If we ever stop because we've hit (b) then we're done (it's important that our shifting and rescaling process preserves what it means to be a progression — a progression in $[N_{r+1}]$ corresponds to a progression in P). So the only issue is if we stop because of case (a).

Crucially, this process can't go on for too long, because every time we run it, the density of our subset increases. More precisely, the density of A_r increases by $\Omega(\delta^k)$ at each step, so the process must stop at time $R \ll \delta^{-k}$ (since the density is bounded above by 1).

We want to analyze how much shorter this makes our interval. It turns out that the dependence of the exponent on δ doesn't really matter, and you can just think of the process as square-rooting the size of N at every step, i.e., $N_{r+1} = N_r^{1/2}$. This means at every step $\log N$ is multiplied by $\frac{1}{2}$, so $\log\log N$ decreases by a constant amount. We have $R \ll \delta^{-k}$ steps, so in order to end up with N not too small, we want to have $\delta^{-k} \ll \log\log N$, which gives the bound stated. (If you plug in the parameters properly, the answer is the same.) □

Remark 2.3. If we could decrease N by a constant factor in (c) instead, then we could remove a log from the final result; but the lemma is not true with this stronger bound. Alternatively, if you could get a *multiplicative* density increment, this would also remove a log.

Now we'll look at why the lemma is true.

§2.2 Quasirandomness

First, we will think of $[N]$ as a subset of $\mathbb{Z}/m\mathbb{Z}$ for a prime $m \in [2N, 4N]$ (this is possible by Bertrand's postulate) — this is useful because having a *finite* group means we can take averages. Crucially, all 4-term progressions in $[N]$ using addition mod m are also 4-term progressions in $[N]$ using normal addition — N is small enough (compared to m) that progressions in $[N] \subseteq \mathbb{Z}/m\mathbb{Z}$ cannot wrap around.

Definition 2.4. The *averaging over progressions* operator Λ is defined as follows: if $f_0, \ldots, f_{k-1}$ are functions $\mathbb{Z}/m\mathbb{Z} \to \mathbb{C}$, then

$$\Lambda(f_0, \ldots, f_{k-1}) := \mathbb{E}_{x,y \in \mathbb{Z}/m\mathbb{Z}}[f_0(x)f_1(x+h)\cdots f_{k-1}(x+(k-1)h)].$$

You can think of Λ as a way of counting progressions in a set, but generalized to functions that don't have to take values in $\{0,1\}$ — in particular,

$$\Lambda(\mathbf{1}_A, \mathbf{1}_A, \cdots, \mathbf{1}_A) = \frac{\#\{k\text{-APs in } A\}}{m^2}$$

(we're counting arithmetic progressions twice — forwards and backwards — but this doesn't really matter).

Notation 2.5. Define $f = \mathbf{1}_A - \delta \mathbf{1}_{[N]}$.

We're essentially taking the indicator function of our set A and subtracting off its density, so that $\mathbb{E}f = 0$. To understand how case (b) can occur, we'll use the following definition:

Definition 2.6. The function f is *ε-quasirandom for progressions* if whenever $f_i = f$ for some i and $|f_i| \leq 1$ for all i, we have

$$|\Lambda(f_0, \ldots, f_{k-1})| \leq \varepsilon.$$

The way to think about this is as the 'random' case of our argument — if A were a random set, we'd expect it to have roughly $\delta^k N^2$ progressions. This definition gives a measure of randomness, with the filter for quasirandomness being by counts of k-term progressions.

Claim 2.7 — If f is ε-quasirandom with $\varepsilon \approx \delta^k$, then case (b) holds.

Proof. We want to find $\Lambda(\mathbf{1}_A, \mathbf{1}_A, \ldots, \mathbf{1}_A)$. We can use a telescoping trick to replace each of the terms with the density of A — we have

$$\Lambda(\mathbf{1}_A, \mathbf{1}_A, \ldots, \mathbf{1}_A) = \Lambda(f, \mathbf{1}_A, \ldots, \mathbf{1}_A) + \Lambda(\delta\mathbf{1}_{[N]}, \mathbf{1}_A, \ldots, \mathbf{1}_A),$$

and since the first term is small we have

$$\Lambda(\mathbf{1}_A, \mathbf{1}_A, \ldots, \mathbf{1}_A) \approx \Lambda(\delta\mathbf{1}_{[N]}, \mathbf{1}_A, \ldots, \mathbf{1}_A).$$

We can repeat this trick to replace the second term and get

$$\Lambda(\mathbf{1}_A, \mathbf{1}_A, \ldots, \mathbf{1}_A) \approx \Lambda(\delta\mathbf{1}_{[N]}, \delta\mathbf{1}_{[N]}, \ldots, \mathbf{1}_A),$$

and so on; eventually we've replaced all the terms, and we have

$$\Lambda(\mathbf{1}_A, \mathbf{1}_A, \ldots, \mathbf{1}_A) \approx \Lambda(\delta\mathbf{1}_{[N]}, \delta\mathbf{1}_{[N]}, \ldots, \delta\mathbf{1}_{[N]})$$

(up to an error of around $k\varepsilon$). But the last term simply counts arithmetic progressions in $[N]$, so it is

$$\frac{\delta^k \#\{k\text{-APs in } [N]\}}{m^2} \asymp \frac{\delta^k N^2}{m^2}.$$

This means $\Lambda(\mathbf{1}_A, \mathbf{1}_A, \ldots, \mathbf{1}_A)$ is close to $\frac{\delta^k N^2}{m^2}$, and therefore the number of k-term arithmetic progressions in A is close to $\delta^k N^2$. □

(The main idea is that if A looks random in the sense of our definition, then for the sake of counting progressions we can essentially just equivocate between A and the entire interval $[N]$ (scaled by the density of A). If we do this k times, then we've reduced the problem to counting progressions in $[N]$, which we know how to do.)

So if f is ε-quasirandom then case (b) holds; we want to see what happens when neither (a) nor (b) holds, so suppose f is not ε-quasirandom.

> **Question 2.8.** If f is *not* ε-quasirandom, then what can we say about A?

There isn't anything obvious we can say — all that we know is that there exist some other functions which make Λ large. But in the case $k = 3$, it turns out that magic happens through Fourier analysis; we'll look at this case first.

§2.3 The $k = 3$ Case

In the case $k = 3$, it turns out that our expression for Λ has a nice equivalent formula on the Fourier side:

$$\mathbb{E}_{x,h} f_0(x) f_1(x+h) f_2(x+2h) = \sum_{\chi} \hat{f}_0(\chi)\hat{f}_1(-2\chi)\hat{f}_2(\chi).$$

The right-hand side can be bounded by $\|\hat{f}_0\|_\infty \cdot \|\hat{f}_1\|_2 \cdot \|\hat{f}_2\|_2$. By Parseval (and the condition on f_1) we have $\|\hat{f}_1\|_2 = \|f_1\|_2 \leq 1$, so we get

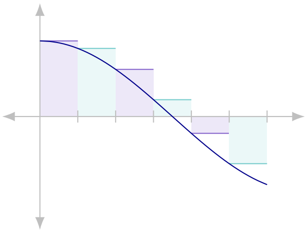
$$\Lambda(f_0, f_1, f_2) \leq \|\hat{f}_0\|_\infty.$$

This means if all the Fourier coefficients of f_0 are small, then $|\Lambda(f_0, f_1, f_2)|$ is necessarily small.

So if f is not ε-quasirandom, then it has a large Fourier coefficient — there exists χ such that $|\hat{f}(\chi)| > \varepsilon$, or in other words

$$\mathbb{E}_x[f(x)e^{-2\pi i\alpha x/m}] > \varepsilon.$$

So we now know that f correlates with some function $e^{2\pi i\alpha x/m}$. It's not obvious how to get from here to (c), which says that A has a density increment on some long progression; but for motivation, we'll consider the special case $\alpha = 1$. In this case, f correlates with the function $e^{2\pi ix/m}$. The nice thing about this function is that it is *slowly varying* — its value on x is basically the same as its value on points near x. This means if we chop up $[N]$ into small intervals of length εN, then $e^{2\pi ix/m}$ is almost constant on each of these pieces — so it's closely approximated by a function that's constant on each piece.



It's a good idea to think of the argument in terms of σ-algebras.

> **Definition 2.9.** A *σ-algebra* $\mathcal{B}$ on $[N]$ is a partition of $[N]$, so that $[N] = \bigcup_{B\in\mathcal{B}} B$.

> **Definition 2.10.** A function $g\colon [N] \to \mathbb{C}$ is *$\mathcal{B}$-measurable* if g is constant on every $B \in \mathcal{B}$.

So if we take $\mathcal{B}$ to be a partition into small intervals, then $e^{2\pi ix/m}$ is very close to being $\mathcal{B}$-measurable.

Definition 2.11. The *conditional expectation* of a function $g\colon [N] \to \mathbb{C}$ is the function

$$\mathbb{E}[g \mid \mathcal{B}]\colon x \mapsto \mathbb{E}_{y \in B_x}[g(y)],$$

where B_x is the cell in our partition that contains x.

The conditional expectation (of an arbitrary function g) is $\mathcal{B}$-measurable, as it is constant on all the cells in the partition. (Essentially, given g we wanted to produce a function that's constant on every cell, and the way we do this is by averaging g over all elements of that cell.)

Our assumption says that f correlates with a function that's almost measurable with respect to a partition in intervals, namely

$$|\langle f, e^{2\pi i x/m}\rangle| \geq \varepsilon.$$

We can replace $e^{2\pi i x/m}$ with a function ϕ that's *actually* constant on each interval (e.g., by setting ϕ on each cell to be the value of $e^{2\pi i x/m}$ at one of its endpoints), so

$$|\langle f, \phi\rangle| \geq \frac{\varepsilon}{2}$$

and ϕ is measurable on a partition of $[N]$ into intervals of size εN.

But now since ϕ is constant on each cell, taking the inner product of f with ϕ is the same as first averaging f on each cell and then taking the inner product of that average with ϕ, so

$$|\langle \mathbb{E}[f \mid \mathcal{B}], \phi\rangle| \geq \frac{\varepsilon}{2}.$$

But we can bound the left-hand side by

$$|\langle \mathbb{E}[f \mid \mathcal{B}], \phi\rangle| \leq \left\lVert \mathbb{E}[f \mid \mathcal{B}] \right\rVert_1 \cdot \left\lVert \phi \right\rVert_\infty .$$

So now we know that $\lVert \mathbb{E}[f \mid \mathcal{B}] \rVert$ is large — in other words, if we divide the world into intervals and average f over each of these intervals, these averages can't cancel too much — f has to have some bias on cells of $\mathcal{B}$ on average. Writing this out explicitly, we have

$$\sum_{b \in \mathcal{B}} \frac{|B|}{N} \left|\mathbb{E}_{x \in B} f(x)\right| \geq \frac{\varepsilon}{2}.$$

This means there must exist some B such that this average is quite large, meaning

$$\left|\mathbb{E}_{x \in B} f(x)\right| \geq \frac{\varepsilon}{2}.$$

But f is just $\mathbf{1}_A - \delta$ on $[N]$, so this means we have

$$\left|\frac{|B \cap A|}{|B|} - \delta\right| \geq \frac{\varepsilon}{2}.$$

So saying that f has bias means that the density of A on these intervals isn't what we'd expect — some intervals must have density too small, and some must have density too large. So we can find some interval B such that the density is too large, for example

$$\frac{|B \cap A|}{|B|} \geq \frac{\varepsilon}{4}.$$

(The inequalities above don't tell us the sign of the deviation, but we know that f has average 0, so if we have lots of big negative deviations then we must also have big positive deviations.)

But this is exactly what we wanted for case (c) (as $\varepsilon \approx \delta^k$) — we now have an interval where A has a density increment, and we can zoom in on this interval and induct.

So far, we've only handled the $\alpha = 1$ case; in general, the function $e^{2\pi i\alpha x/m}$ is *not* slowly varying (and instead oscillates a lot), so splitting $[N]$ into intervals won't work — this function is no longer almost $\mathcal{B}$-measurable if $\mathcal{B}$ is a partition into intervals. But we don't exactly need to split $[N]$ into intervals — all we needed was to split it into *progressions*. And for any α, we can find some spacing t such that if we only sample from $\{x, x+t, x+2t, \ldots\}$ then $e^{2\pi i\alpha x/m}$ looks like it's slowly varying — so if we chop the world up into progressions with this common difference, then our function again looks almost constant and the same argument works. (The fact that we can choose such a spacing isn't totally obvious; you essentially want to choose a shift so that $\frac{\alpha t}{m}$ is small mod 1, so that hopping along the shift doesn't change the phase that much. But this is possible.)

§2.4 The $k = 4$ Case

We'll now try to see that an idea like this still works when $k = 4$ (even though the Fourier analysis won't work anymore). The first idea is the following inequality (shown by Gowers):

Lemma 2.12

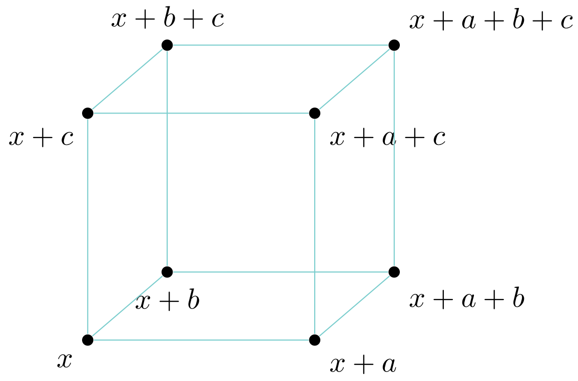
The quantity $|\Lambda(f_0, f_1, f_2, f_3)| = |\mathbb{E}_{x,h}[f_0(x)f_1(x+h)f_2(x+2h)f_3(x+3h)]|$ is bounded above by

$$\left(\mathbb{E}_{x,a,b,c} f_0(x)\overline{f_0(x+a)f_0(x+b)}f_0(x+a+b)\overline{f_0(x+c)}f_0(x+a+c)f_0(x+b+c)\overline{f_0(x+a+b+c)}\right)^{1/8}.$$

Definition 2.13. The *Gowers U^3 norm of f*, denoted $\|f\|_{U^3}$, is the above quantity

$$\left(\mathbb{E}_{x,a,b,c} f(x)\overline{f(x+a)f(x+b)}f(x+a+b)\overline{f(x+c)}f(x+a+c)f(x+b+c)\overline{f(x+a+b+c)}\right)^{1/8}.$$

Essentially, $\|f_0\|_{U^3}$ averages f_0 over things that look like cubes, with suitable complex conjugate signs.



This can be proven by applying the Cauchy–Schwarz identity three times. The same is also true if we replace f_0 with the functions in any of the other positions.

So now the point is that if $|f|_{U^3} \le \varepsilon$, then f is ε-quasirandom for progressions. So it suffices to study the case where $|f|_{U^3}$ is big. If we can say something about such functions f — namely, that they have some bias on progressions — then that provides the missing piece for our argument.

Question 2.14. What can we say about f if $\|f\|_{U^3}$ is large?

Example 2.15

If $f(x) = e^{2\pi i \alpha x^2/m}$, then $\|f\|_{U^3}$ is 1 (which is as big as it can possibly be) — this is because

$$x^2 - (x+a)^2 - (x+b)^2 + \cdots - (x+a+b+c)^2 = 0$$

(we can think of this as differentiating x^2 three times), so we're simply averaging $e^0 = 1$.

The same is true if we replace x^2 with any quadratic. So these quadratic functions are sort of the spirit animal of functions with large U^3 norm — quadratic functions have very large U^3 norm, and we'd like to say that any function with large U^3 norm is trying to be quadratic in some sense.

In the case $k = 3$, we saw that if f fails to be quasirandom, then it *correlates* with some function $e^{2\pi i \alpha x/m}$. We would *like* something similar to be true in the case $k = 4$ — it would be nice if we could say that

$$|\langle f, \phi \rangle| \gg_\varepsilon 1$$

for some function $\phi(x) = e^{2\pi i q(x)/m}$, since then (similarly to before) we could turn this quadratic bias into a bias on short progressions, which would be great. Unfortunately, this is false — there exist functions with large U^3 norm which don't correlate with these quadratics.

But Gowers found a way to still make this work — the point is that if we split the world into progressions, and then we split those progressions into smaller progressions, we've still split the world into progressions. Gowers noted that it's too much to hope that on the *whole* interval, our function correlates with something quadratic. But maybe if we first break up our interval into smaller pieces, we can arrange that on each of these short progressions the function *does* correlate with a quadratic function. This is a weaker statement since it says nothing about how these quadratics in different parts relate to each other — so the quadratic functions on the short progressions don't necessarily expand to one on the whole group — and it turns out to be true.

Theorem 2.16 (Gowers)

If $f\colon \mathbb{Z}/m\mathbb{Z}$ satisfies $\|f\|_\infty \leq 1$ and $\|f\|_{U^3} \geq \varepsilon$, then there exists a partition $\mathcal{B}$ of $\mathbb{Z}/m\mathbb{Z}$ into progressions of length $N^{\varepsilon^{O(1)}}$ and a function $\phi\colon \mathbb{Z}/m\mathbb{Z} \to \mathbb{C}$ with $\|\phi\|_\infty \leq 1$, such that

$$|\langle f, \phi \rangle| \gg \varepsilon^{O(1)}$$

and $\phi|_B$ is a quadratic function for each $B \in \mathcal{B}$ (i.e., if we parametrize B as $y_0 + th$, then we should have $\phi(y_0 + th) = e^{2\pi i q(t)/m}$ for some quadratic polynomial q).

Then using this, we can find some $B \in \mathcal{B}$ which the correlation favors, and zoom in on it. Then we need to further shatter that B into progressions that make the quadratic phase roughly constant, and zoom in on one of these as before.

§3 Functions with Large U^3 Norm

The above result is exactly what Gowers needed for the proof of Szemerédi's theorem. But we might expect that something stronger is true. This theorem tells us that ϕ is quadratic on each of our pieces B individually, but it doesn't tell us anything about how these pieces relate to each other. But more than this should be true — it should be possible to make a global statement. As mentioned earlier, it's not true that f correlates to a quadratic function, but there *is* a stronger statement that's true — where rather than dividing the world into tiny progressions, we consider a *multidimensional progression*.

Definition 3.1. A *d-dimensional progression* is a set $Q = \{a_0 + a_1t_1 + \cdots + a_dt_d\}$ where $a_i \in \mathbb{Z}/m\mathbb{Z}$ are fixed, and t_i are integers ranging over some intervals $[T_i]$.

(You can think of a d-dimensional progression as taking a box in d dimensions and projecting it down; the a_i correspond to the different directions.)

Theorem 3.2

We can find a d-dimensional progression Q with $d \ll_\varepsilon 1$ and $|T_1| \cdots |T_d| \gg_\varepsilon m$ and a quadratic polynomial $q\colon \mathbb{Z}^d \to \mathbb{R}$ such that $|\langle f, \phi\rangle| \gg 1$, where

$$\phi(x) = \mathbf{1}_Q(x) \cdot e^{2\pi i q(t_1,\dots,t_d)}$$

(where if $x \in Q$ the t_i are the ones for which $x = a_0 + a_1t_1 + \cdots + a_dt_d$).

This gives a sort of global statement — we can essentially find a huge box (almost as big as the entire group) and a quadratic polynomial on that box which f correlates with. (Note that now our quadratic polynomial is in d variables, so we can square these variables *and* take their products — there are a lot more d-variable quadratic polynomials than 1-variable ones.)

Remark 3.3. If the sets $|T_i|$ become too big, then every point in $\mathbb{Z}/m\mathbb{Z}$ ends up inside the box, but there is ambiguity about which t_i correspond to a point x (as the above definition assumes that they're unique, and if the box is too large this will not be the case). In that case, you can average over the different lifts t_i, which does give a related function that's defined everywhere. So you can either choose a small box that's tightly controlled, or a huge box that maps down to everything but has ambiguity you need to resolve.

§3.1 Ideas Behind the Proof

We'll now briefly sketch how you can start trying to prove a result like this.

It's useful to keep in mind the motivating example where f really is quadratic. If we take two derivatives of the quadratic, by considering

$$f(x)\overline{f(x+a)f(x+b)}f(x+a+b),$$

the result is not 0, but it *is* constant — we will get $e^{2\pi i(2\alpha ab)/m}$, which notably doesn't depend on x and is bilinear as a function of a and b.

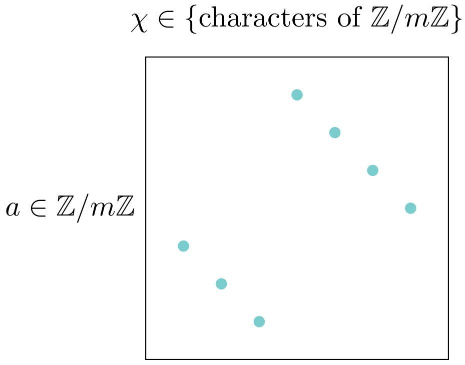
In general (no longer assuming f is quadratic), the above expression won't be independent of x, but we can *make* it independent of x by averaging.

Definition 3.4. Define the function $Sf(a,b) := \mathbb{E}_x f(x)\overline{f(x+a)f(x+b)}f(x+a+b)$.

In the special case where f is quadratic, $Sf(a,b)$ is bilinear in a and b (or rather, its exponent is). If we now imagine fixing a and thinking of this as a function of b, then the exponent is just a linear function of b — so we have an expression of the form $e^{2\pi i(-)b}$. So if we take the Fourier transform, then we just get a single point — we have

$$\mathbb{E}_b Sf(a,b)\chi(-b) = \begin{cases} 1 & \text{if } \chi = e^{2\pi a(-)/m} \\ 0 & \text{otherwise.} \end{cases}$$

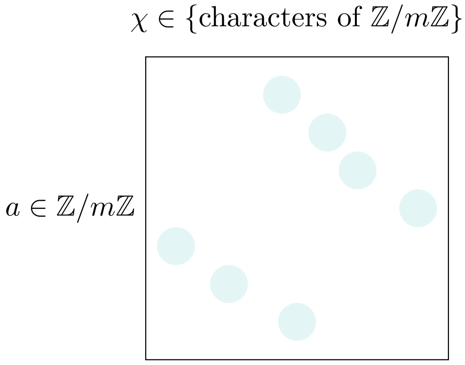
If we imagine a box with $\mathbb{Z}/m\mathbb{Z}$ on one axis (corresponding to a) and the characters of $\mathbb{Z}/m\mathbb{Z}$ on the other (corresponding to χ), and we plot the points at which this expression is 1, then the result is the graph of a linear function (from $\mathbb{Z}/m\mathbb{Z}$ to its characters).


$\chi \in \{\text{characters of } \mathbb{Z}/m\mathbb{Z}\}$
$a \in \mathbb{Z}/m\mathbb{Z}$

In the general case, we can still consider the Fourier transform

$$\hat{S}f(a,\chi) = \mathbb{E}_b Sf(a,b)\chi(-b).$$

This is no longer going to look like a perfect graph of a function — it will be fuzzy — but it will have some nice properties. We want to say that in some sense, it looks a little bit like the graph of a function that is a little bit linear.


$\chi \in \{\text{characters of } \mathbb{Z}/m\mathbb{Z}\}$
$a \in \mathbb{Z}/m\mathbb{Z}$

More precisely, here are some properties:

- We have $\hat{S}f(a,\chi) \geq 0$ for all (a,χ) — so we can think of it as a probabilistic function mapping out a *distribution* of possible outputs (rather than just a single number, as in the quadratic case).
- If we fix a and sum over the weights of each χ, then $\sum_\chi \hat{S}f(a,\chi) \leq 1$. (This also corresponds to the probabilistic function interpretation.)
- We have $\mathbb{E}_a \sum_\chi \hat{S}f(a,\chi)^2 \geq \varepsilon^8$ — so if $|f|_{U^3} \geq \varepsilon$ then this object *is* a bit concentrated (in other words, it looks a little like a graph).
- Imagine we go through our 'fuzzy function' and make a choice for every a to turn it into an actual function τ. We can choose $\tau(a)$ for each a such that $\hat{S}f(a,\tau(a)) \geq \varepsilon^{10}$ (i.e., our fuzzy function has a reasonable amount of weight on $(a,\tau(a))$) and such that τ is a 'little bit linear' — more precisely, the number of (a,b,c) such that

  $$\tau(a) - \tau(a+b) - \tau(a+c) + \tau(a+b+c) = 0$$

  is at least $\varepsilon^{10}N^3$. (If τ were actually linear, then this would always be true; here it's true a reasonable fraction of the time.)

Then you can apply a bunch of theorems in additive combinatorics (this involves finding a piece with small doubling and covering it with progressions), to end up with a nice multi-dimensional progression that smothers all the large points of your function. Then you can squash that multidimensional progression back onto the a-axis, giving the d-dimensional progression; and the function taking a and spitting out $\alpha a\tau(a)$ or something similar will correspond to the quadratic function.

Remark 3.5. You can also consider $A \subseteq [N]^2$ (i.e., we're considering a box in $\mathbb{R}^2$) with $|A| \geq \delta N^2$, in which we want to find a square. (This is the first case of the multidimensional Szemerédi theorem.) We don't know sensible bounds for this (better than tower-type) — the problem is that even a weak Gowers inverse theorem in this case isn't known, and we don't even know what the right statement is. The problem is that structured quadratics $e^{q(x,y)}$ are bad, but so are *any* functions of the form $g(x)h(y)$ — this makes it hard to write down a concise statement that captures everything.