The quantitative ergodic theory proof of Szemerédi's theorem

Peter Rudzis

6 June 2016

Contents

1 Introduction

A subset S of positive integers is said to have *positive upper density* if

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

An intriguing fact, first conjectured by Erdös and Turàn [3] in 1936 and proven by Szemerédi [12] in 1975, is that every subset S of positive integers with positive upper density contains arbitrarily long arithmetic progressions.[1] It is surprising that such a simple condition on the density of a set of integers is enough to guarantee such lockstep regularity. This fact can be equivalently phrased in terms of finite sets, and doing so will turn out to be more useful.

Theorem 1.1 (Szemerédi's theorem)**.** *Given an integer $k \geq 0$ and a density $0 < \delta \leq 1$, there is an integer $N_{SZ}(k,\delta)$ such that for every $N \geq N_{SZ}(k,\delta)$, every set $A \subset \{1,...,N\}$ of density $|A|/N \geq \delta$ contains an arithmetic progression of length k.*

By investigating this and related questions, Erdös and Turàn originally hoped to find a good explicit bound for $N_{SZ}(k,\delta)$ [7]. If $N_{SZ}(k,\delta) \leq 2^{c_k\delta^{-1}}$ where c_k is a constant depending on k, this would imply the following remarkable theorem, recently proven by Ben Green and Terence Tao in 2007 [7].

Theorem 1.2. *The prime numbers contain infinitely many arithmetic progressions of length k, for all positive integers k.*

In fact, mathematicians have yet to find a bound nearly as good as $N_{SZ}(k,\delta) \leq 2^{c_k\delta^{-1}}$. Green and Tao do not need the explicit bound for their proof of arbitrarily long arithmetic progressions in the primes, although Szemerédi's theorem nonetheless plays a fundamental role.

Szemerédi's theorem becomes non-trivial to prove when $k \geq 3$. (For $k = 1$ or $k = 2$, just take $N \geq N_{SZ}(k,\delta) = \lceil k/\delta \rceil$, and observe that every subset of $\{1,...,N\}$ of density at least δ must contain at least k elements.) Progress was first made by Roth [9] in 1953, by proving the case where $k = 3$, using methods from Fourier analysis. The general case was finally proven in 1975 by Szemerédi [13], by way of an intricate combinatorial argument. Since then other proofs have been given. The one which has garnered the most attention is that of Furstenberg [4] in 1977, who famously drew upon the methods of ergodic theory to make his argument. The ergodic theory proof has led to a greater level of interaction between the disciplines of ergodic theory, number theory, and combinatorics, which continues to this day [2][8]. Another proof worth mentioning is that of Gowers [6] in 1998, which extends the Fourier analysis approach of Roth to the general case and which achieved the best known bound so far: $N_{SZ}(k,\delta) \leq 2^{2^{\delta^{-c_k}}}$.

[1] By the length of an arithmetic progression, we mean the number of terms in the sequence. Warning: this is *not* a claim that S contains an infinite arithmetic progression.

Terence Tao views Furstenberg's ergodic theory argument as the simplest in the sense that deep results from ergodic theory allow for a relatively short proof. On the other hand, the argument suffers for the same reason, requiring the reader to have a knowledge of ergodic theory to understand the proof. Tao's quantitative ergodic theory proof [14] of Szemerédi's theorem takes Furstenberg's argument as its inspiration. However, the more abstract setting of a measure preserving probability system is replaced by the Hilbert space $L^2(\mathbb{Z}_N)$ of functions $f : \mathbb{Z}_N \to \mathbb{C}$, where $\mathbb{Z}_N$ denotes the cyclic group of integers in addition modulo N. As a consequence, the arguments are more elementary, making for a proof which, with the exception of a few basic results, is self-contained and which is accessible to a broader audience.

Tao observes in [14] that all the proofs of Szemerédi's theorem are achieved by establishing a kind of *dichotomy between randomness and structure*. He gives the following heuristic outline for their construction.

1. Given a set A, establish some proxy for A (e.g. a set of maps defined on A).
2. Define a concept of *randomness* and concept of *structure* in the proxy for A.
3. Establish a *structure theorem* which splits the proxy into random and structured components.
4. Establish a type of *generalized Von Neumann theorem* to eliminate the random component and a type of *structured recurrence theorem* to extract the structured component.[2]

Among the more fascinating aspects of Tao's quantitative ergodic theory proof is his way of approaching steps 2 and 3, to establish a dichotomy between randomness and structure. The notion of randomness in subsets of $L^2(\mathbb{Z}_N)$ is quantified by *Gowers uniformity norms*, first introduced in [6]. Functions lacking *quasiperiodic* structure will have small Gowers uniformity norms. Structure on the other hand is quantified by *uniform almost periodicity norms*. Once these norms have been introduced, the structure theorem is proven via an intricate *energy incrementation argument*.

This paper presents some of the key ideas and methods used in Tao's quantitative ergodic theory proof [14].[3] Pursuing the heuristic outline above, this paper is organized as follows: Section §2 gives a brief introduction to some of the main concepts from Tao's paper, including the the Hilbert space of functions $L^2(\mathbb{Z}_N)$, which serves as a proxy for the set of integers $\{1, ..., N\}$, and Szemerédi's theorem is rephrased as a finitary ergodic theory problem. Next presented are Tao's versions of the structure theorem, generalized Von Neumann theorem, and structured

[2] These names come from their respective ergodic theory analogs, the *Von Neumann mean ergodic theorem* and the *Poincaré recurrence theorem* respectively. We do not discuss these theorems in this paper. For more information about these theorems, see [2].

[3] From this point forward, the results we present can be assumed to have come from [14] unless indicated otherwise.

recurrence theorem which form the backbone of the argument. It is shown how they fit together to prove Szemerédi's theorem. Section §3 formally defines *Gowers uniformity norms* and *uniform almost periodicity norms* and explains their duel relationship. Tao suggests the possibility of an alternative proof for the generalized Von Neumann theorem, which uses properties of uniform almost periodicity norms, and we work out the details of this proof in §4. Section §5 presents some of the details of Tao's *energy incrementation argument* for proving the structure theorem.

2 Overview of the proof

2.1 The $\mathbb{Z}_N \to \mathbb{C}$ setting: definitions and notation for Tao's proof

Let $\mathbb{Z}_N := \mathbb{Z}/N\mathbb{Z}$ denote the cyclic group of integers modulo N. For the purposes of proving theorem 1.1, we are interested in the subsets of $\mathbb{Z}_N$ of density δ. However, following the outline given in the introduction, we will not examine these sets directly but instead via a proxy for $\mathbb{Z}_N$. For describing this proxy Tao draws on ideas and notation from probability theory and ergodic theory.

Notation 2.1. If S is a finite set, and f is a (real or complex valued) function defined on S, then the *expected value* of f on S is denoted in several different ways.

$$\frac{1}{|S|}\sum_{x\in S} f(x) =: \mathbb{E}_{x\in S} f(x) = \mathbb{E}_S f = \int_{x\in S} f(x) = \int_S f.$$

We will also denote the *indicator*, or characteristic, function by

$$\mathbf{1}_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases},$$

where $\mathbf{1}_A$ is defined on S and $A \subset S$.

As a general rule, the integral notation for the expected value is favored if it is desirable to emphasize expected value as being a kind of *measure*[4] or when it is easier on the eyes. We now define formally the proxy to be used for $\mathbb{Z}_N$.

Definition 2.1 (The $L^2(\mathbb{Z}_N)$ Hilbert space)**.** By $L^2(\mathbb{Z}_N)$ we denote the vector space of functions $f : \mathbb{Z}_N \to \mathbb{C}$, equipped with inner product

$$\langle f, g\rangle := \int_{\mathbb{Z}_N} f\overline{g} \tag{1}$$

and induced norm

$$||f||_{L^2} := \langle f, f\rangle^{1/2} = \left(\int_{\mathbb{Z}_N} |f|^2\right)^{1/2}. \tag{2}$$

[4]The concept of *measure* generalizes ideas such as integration, probability, and other ways to assign a size to a set. It is not needed to understand what is discussed in this paper.

Although we will not prove it here, it is not too difficult to show that $L^2(\mathbb{Z}_N)$ is truly a normed vector space over the complex numbers. In fact, $L^2(\mathbb{Z}_N)$ is even better. We record the following properties of $L^2(\mathbb{Z}_N)$.

1. $L^2(\mathbb{Z}_N)$ is a vector space over $\mathbb{C}$.
2. The inner product $\langle \cdot, \cdot \rangle$ is linear in each argument (bilinearity) and satisfies

   $$\langle f, g \rangle = \overline{\langle g, f \rangle} \quad \text{and} \quad \langle f, f \rangle \geq 0$$

   for all $f, g \in L^2(\mathbb{Z}_N)$.
3. $||f||_{L^2} = 0$ if and only if $f = 0$.
4. Both the Cauchy-Schwartz and the Triangle inequalities hold

   $$|\langle f, g \rangle| \leq ||f||_{L^2} ||g||_{L^2} \quad \text{and} \quad ||f + g||_{L^2} \leq ||f||_{L^2} + ||g||_{L^2}.$$

5. $L^2(\mathbb{Z}_N)$ is *complete*. That is, all Cauchy sequences in $L^2(\mathbb{Z}_N)$ converge.
6. $L^2(\mathbb{Z}_N)$ is *separable*. That is, $L^2(\mathbb{Z}_N)$ contains a subset which is countable and dense.[5].

More generally, any space equipped with norm and induced inner product that satisfies the above six properties is known as a *Hilbert space* [11].

When discussing $L^2(\mathbb{Z}_N)$, we will almost always want to take N to be prime, and we adopt the convention that

Notation 2.2. *N is a large prime number.*

The purpose for this comes from the following basic fact from number theory.

Proposition 2.1. *If N is prime and nonzero integer $n < N$, then the map $m \mapsto mn(modN)$ takes $\mathbb{Z}_N$ bijectively to itself. [8]*

This additional multiplicative structure will prove useful in at least several of the arguments to follow.

We establish two other pieces of notation.

Notation 2.3. For each $f \in \mathbb{Z}_N$, the supremum norm will be denoted by

$$||f||_{L^\infty} := \sup_{x \in \mathbb{Z}_N} |f(x)|$$

Notation 2.4. We write $f = O_{p_1,...,p_m}(g)$ if $|f(x)| \leq \alpha g(x)$ for some fixed α, possibly depending on parameters $p_1, ..., p_m$.

[5]For an analogue in real vector space $\mathbb{R}^n$, consider the subset of rational points $\mathbb{Q}^n$ which is countable and dense in $\mathbb{R}^n$

Lastly we adopt the following curious notion of *boundedness*.

Definition 2.2 (boundedness). We say that a function $f : \mathbb{Z}_N \to \mathbb{C}$ is *bounded* if $||f||_{L^\infty} \leq 1$.

This definition of boundedness is in fact the natural one to adopt for the purpose of interpreting sizes of functions in $L^2(\mathbb{Z}_N)$ probabilistically (since probabilities never exceed 1). For example, the expected value of the indicator function $\mathbf{1}_A$ over $\mathbb{Z}_N$, where $A \subset \mathbb{Z}_N$, is precisely the probability of picking an element of A at random from $\mathbb{Z}_N$.

2.2 Rephrasing Szemerédi's theorem as a quantitative ergodic theory problem

Suppose we have a subset $A \subset \mathbb{Z}_N$ and we are looking for an arithmetic progression $x, x+r, ..., x+(k-1)r \in A$. The natural way to transfer this query to the proxy $L^2(\mathbb{Z}_N)$ is to look for periodic behavior in the indicator function $\mathbf{1}_A \in L^2(\mathbb{Z}_N)$. That is, we want an x, r, and k such that $\mathbf{1}_A(x) = \mathbf{1}_A(x+r) = \cdots = \mathbf{1}_A(x+(k-1)r) = 1$. With this motivation we define the shift map.

Definition 2.3 (Shift map). *For each integer n, the shift map $T^n : L^2(\mathbb{Z}_N) \to L^2(\mathbb{Z}_N)$, is defined by*

$$T^n(f)(x) := f(x+n), \quad x \in \mathbb{Z}_N.$$

The shift map enjoys some nice properties. Two of them, which are easy to see from its definition, are that it distributes over addition and multiplication.

$$T^n f + T^n g = T^n(f+g), \quad (T^n f)(T^n g) = T^n(fg) \quad \text{for} \quad f, g \in L^2(\mathbb{Z}_N)$$

Another property which the shift map has in $L^2(\mathbb{Z}_N)$ is that it preserves expectation. In particular,

Proposition 2.2. *If $f \in L^2(\mathbb{Z}_N)$, then*

$$\mathbb{E}_{Z_N} f = \mathbb{E}_{Z_N} T^n f$$

for any $n \in \mathbb{Z}$.

This result follows quickly from the observation that, under addition modulo N, shifting the arguments of f by n simply permutes the terms of the sum, and does not change the expected value.

Saying that A contains an arithmetic progression of length k is the same as saying that the product $\mathbf{1}_A(x)T^r\mathbf{1}_A(x)T^{2r}\mathbf{1}_A(x)\cdots T^{(k-1)r}\mathbf{1}_A(x)$ does not equal zero for some x and some r. This motivates the following restatement of Szemerédi's theorem in terms of shift maps.

Theorem 2.1 (Finitary ergodic theory statement of Szemerédi's theorem)**.** *For any integer $k \geq 1$, any large prime integer $N \geq 1$, any $0 < \delta \leq 1$, and any non-negative bounded function $f : \mathbb{Z}_N \to \mathbb{R}^+$ with*

$$\int_{\mathbb{Z}_N} f \geq \delta \tag{3}$$

we have

$$\mathbb{E}_{r\in\mathbb{Z}_N} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} f \geq c(k, \delta) \tag{4}$$

for some $c(k, \delta) > 0$.

Proof theorem 2.1 implies theorem 1.1. Fix k, δ and let $N \geq 0$ be a large integer to be chosen later. Suppose $A \subset \{1, ..., N\}$ such that $|A| \geq \delta N$. By Bertrand's postulate[6] there is a prime number $N' \in [kN, 2kN]$. Define A' to be the image of A under the inclusion map from $\{1, ..., N\}$ to $\mathbb{Z}_{N'}$ (i.e. we have the set equality $A' = A$, but A' is equipped with the operation of addition modulo N'.) Observe that $\int_{\mathbb{Z}_{N'}} \mathbf{1}_{A'} = |A'|/N' \geq \delta/2k$. Thus, by theorem 2.1,

$$\mathbb{E}_{r\in\mathbb{Z}_N} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} \mathbf{1}_{A'} \geq c(k, \delta/2k) \tag{5}$$

Note that the product $\prod_{j=0}^{k-1} T^{jr}\mathbf{1}_{A'}(x)$ is equal to 1 if all the shifts $x, x - r, ..., x - (k-1)r$ are contained in A' and the product is equal to 0 otherwise. Thus, expanding the the expectation operators, we get

$$|\{(x, r)|x, x + r, ..., x + (k - 1)r \in A'\}| \geq c(k, \delta/2k)(N')^2 \geq c(k, \delta/2k)k^2N^2$$

It is clear that the number of points of the form $(x, 0)$ in the set on the left is exactly $|A'|$ (we have one point for each $x \in A'$). Thus we get the inequality

$$\begin{aligned} |\{(x, r)|r \neq 0, \text{ and } x, x + r, ..., x + (k - 1)r \in A'\}| &\geq c(k, \delta/2k)k^2N^2 - |A'| \\ &\geq c(k, \delta/2k)k^2N^2 - N \end{aligned}$$

If we choose $N > 1/c(k, \delta/2k)k^2$, then the left hand side of the above inequality will be positive, and we conclude that there is a progression $x, x+r, ..., x+(k-1)r$ in A'. We are not quite done, since this is a progression in addition modulo N'. However, since $x \in A'$, we have $0 < x \leq N$ and since $x + r \in A'$, we have $-N \leq r \leq N$. Since $N' > kN$, under ordinary integer addition the arithmetic progression $x, x + r, ..., x + (k - 1)r$ is contained in the $\{1, ..., N'\}$, and by the equality of the sets A and A', the arithmetic progression is contained in A, as desired. □

Remark 2.1. Theorem 2.1 is in fact equivalent to theorem 1.1. According to Tao, the reverse implication can be proven easily from some work by Varnavides [15].

[6] For any positive integer n, there is a prime between n and $2n$. (See [14])

2.3 The dichotomy of randomness and structure

As mentioned in the introduction, finding a way to quantify appropriate notions of randomness and structure in subsets of $\mathbb{Z}_N$ is an important part of Tao's quantitative ergodic theory proof. Randomness is quantified by Gowers uniformity norms $||\cdot||_{U^d}$, $d \in \mathbb{Z}^+$, first described in [6]. In particular, functions which exhibit a lack of periodic (i.e. random) behavior will have small Gowers uniformity norms.

Structure, on the other hand, is quantified by *uniform almost periodicity norms*. These norms are defined on the space of *uniform almost periodic functions*, Tao's own innovation, and act as a kind of duel to Gowers uniformity norms. As a heuristic, Tao describes a uniform almost periodic function $f : \mathbb{Z}_N \to \mathbb{C}$ of order $k-2$ as being a function which behaves like

$$F(x) := \frac{1}{J}\sum_{j=1}^{J} c_j e^{2\pi i P_j(x)/N}, \tag{6}$$

where $x \in \mathbb{Z}_N$, $c_j \in \mathbb{C}$ is a constant with $|c_j| \leq 1$, and P_j is a polynomial of order at most $k-2$ with coefficients in $\mathbb{Z}_N$. In keeping with Tao's terminology, such a function shall be referred to informally as a *quasiperiodic periodic function of order* $k-2$. Gowers uniform functions and uniform almost periodic functions are, in a sense, orthogonal in $L^2(\mathbb{Z}_N)$.

Defining formally both the Gowers uniformity norms and the uniform almost periodicity norms takes a somewhat lengthy setup. For this reason, we postpone this discussion and choose to first describe the three theorems which form the backbone of Tao's argument.[7] In doing so, we will not need to understand all the properties of the Gowers uniformity and uniform almost periodicity norms. We will however then have better motivation for defining each norm formally in §3.

2.4 Three important subtheorems

Here we look at Tao's proof from a birds eye point of view, describing the three major subtheorems which together imply theorem 2.1. The first of these theorems says that the random (i.e. *Gowers uniform*) error, which will arise when estimating (4), is of negligible size.

Theorem 2.2 (Generalized Von Neumann theorem)**.** *If $k \geq 2$ and $\lambda_0, ..., \lambda_{k-1}$ are distinct elements of $\mathbb{Z}_N$, then for any bounded functions $f_0, ..., f_{k-1} : \mathbb{Z}_N \to \mathbb{C}$, we have*

$$\left| \mathbb{E}_{r\in\mathbb{Z}_N} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{\lambda_j r} f_j \right| \leq \min_{0\leq j\leq k-1} ||f_j||_{U^{k-1}}. \tag{7}$$

[7] This is also the same order of exposition as given by Tao in [14].

Of the three theorems in this section, the proof of theorem 2.2 is the easiest. Tao's proof relies on some basic facts about Gowers uniformity norms, and he uses induction on k. Tao mentions that it is also possible to give a proof of the theorem using certain properties of the uniform almost periodicity norms. We work out the details of this alternative proof in §4.

Next we state theorems 2.3 and 2.4, which are in some ways complementary. The statements of each theorem may be confusing due to the number of parameters floating about to keep track of. For this reason we have compiled a list of the main actors in each theorem.

1. k and δ – respectively, the length of an arithmetic progression and the density of a subset of $\mathbb{Z}_N$.

2. $f_{U^\perp}$ – an "anti-Gowers uniform" function. That is, if a function f_U has a small Gowers uniformity norm than we have the orthogonality condition that $\langle f_U, f_{U^\perp}\rangle$ is small.

3. f_{UAP} – a uniform almost periodic function.

4. d – the "order" of uniform almost periodicity of f_{UAP} (to be defined in the next section). Eventually we want $d = k - 2$. However, the theorem is stated for a general d for purposes of induction.

5. M – a control on the size of a uniform almost periodic function, which prevents it from blowing up to infinity.

6. N_1 and μ – two (somewhat contrived) parameters to be used in the inductive proof of theorem 2.3. Ultimately, we want $N_1 = N - 1$ and $\mu = 1$.

7. $F(x)$ – a real-valued function to control the size of the "random error" which arises when computing (4).

The first theorem states that if we are working with a "Gowers anti-uniform" function $f_{U^\perp}$ – in this case, a function which is close to a uniform almost periodic function – then we can place an estimate on a kind of average value on the products of shifts of $f_{U^\perp}$, quite similar to (4).

Theorem 2.3 (Almost periodic functions are recurrent theorem)**.** *Suppose $d \geq 0$ and $k \geq 1$ are integers, and $f_{U^\perp}$, f_{UAP} are real-valued, non-negative, bounded functions satisfying*

$$||f_{U^\perp} - f_{UAP}||_{L^2} \leq \frac{\delta^2}{1024k} \tag{8}$$

$$\int_{\mathbb{Z}_N} f_{U^\perp} \geq \delta \tag{9}$$

$$||f_{UAP}||_{UAP^d} < M \tag{10}$$

for some $0 < \delta, M < \infty$. Then for all $\mu \in \mathbb{Z}_N$ and $N_1 > 0$, we have

$$\mathbb{E}_{0<r<N_1} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{\mu jr} f_{U^\perp} \geq c_0(d, k, \delta, M)$$

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

The next theorem says that we can split any real valued non-negative function f into a random (Gowers uniform) part and structured (Gowers anti-uniform) part.

Theorem 2.4 (Structure theorem)**.** *Suppose $k \geq 3$ and f is a real-valued, non-negative function which obeys $\int_{\mathbb{Z}_N} f \geq \delta$. Pick an arbitrary function $F : \mathbb{Z}_N \to \mathbb{R}^+$. Then we can find positive number $M = O_{k,\delta}(1)$, function f_U bounded, and functions $f_{U^\perp}$ and f_{UAP} non-negative and bounded, such that we may split*

$$f = f_U + f_{U^\perp}$$

and (8),(9), and (10) hold true for $d = k - 2$ and moreover

$$||f_U||_{U^{k-1}} \leq \frac{1}{F(M)} \tag{11}$$

Theorem 2.4 – the structure theorem – is the easier than theorem 2.3 (which is not to say easy). Both of the theorems invoke an energy incrementation argument, which in each case reduces the proof to a matter of proving a dichotomy. Section §4 gives a definition for energy, states some results about energy, and describes its relationship to the Gowers uniform and uniform almost periodic functions. With this setup, the detailed proof of theorem 2.3, via the energy incrementation argument is then presented.

We conclude this section by giving the proof that the three theorems in this section prove theorem 2.1.

Proof theorems 2.2,2.3,2.4 imply theorem 2.1. Choose integer $k \geq 1$, large prime number $N \geq 1$, real number $0 < \delta \leq 1$, and non-negative bounded function $f : \mathbb{Z}_N \to \mathbb{R}^+$. Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be a function to be chosen later. We have already discussed the trivial cases of $k = 1$ and $k = 2$ in the introduction, so in this proof we can take $k \geq 3$. By the Structure Theorem (theorem 2.4) we can find an $M = O_{k,\delta}(1)$, a bounded function f, and bounded non-negative functions $f_{U^\perp}$ and f_{UAP} such that $f = f_{U^\perp} + f_{UAP}$ and estimates (8), (9), (10), and (11) hold. For the calculation to follow, it is useful to define the set S_k to be the set of all maps $s : \{0, ..., k-1\} \to \{f_{U^\perp}, f_U\}$ and the set $S_k^- := S_k - \{s_0(j) := f_{U^\perp}\}$. We have

$$\mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} f = \mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} (T^{jr} f_{U^\perp} + T^{jr} f_{UAP})$$

$$= \mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \sum_{s\in S_k} \prod_{j=0}^{k-1} T^{jr} s(j) = \sum_{s\in S_k} \mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} s(j)$$

$$= \mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} f_{U^\perp} + \sum_{s\in S_k^-} \mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} s(j) \tag{12}$$

Taking $N_1 = N$, $\mu = 1$, and $d = k - 2$ in theorem 2.3, we have the estimate

$$\mathbb{E}_{r\in\mathbb{Z}_N} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} f_{U^\perp} \geq c_0(k-2, k, \delta, M)$$

By the generalized Von Neumann theorem (theorem 2.2) and the estimate (11) for $||f_U||_{U^{k-1}}$ in theorem 2.3, we also have

$$\mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} s(j) \leq \min\{||f_{U^\perp}||_{U^{k-1}}, ||f_U||_{U^{k-1}}\} \leq \frac{1}{F(M)}$$

(We were able to drop the $|\cdot|$ signs in the above estimate because each of the $s(j)$ equals either f_U or $f_{U^\perp}$ which are non-negative and real.) Note that there are $2^k - 1$ elements in S_k^-, so comparing these estimates with (13) yields

$$\mathbb{E}_{r\in\mathbb{Z}_\mathbb{N}} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{jr} f \geq c_0(k-2, k, \delta, M) - \frac{2^k - 1}{F(M)} \tag{13}$$

We are free to choose the value of $F(M)$ large enough that the right side of (13) is positive. Since $M = O_{k,\delta}(1)$, the right side of the (13) is equal to some positive quantity depending only on k and δ, which is what we needed to show. □

3 Being precise about norms

Here we present the formal definitions of the Gowers uniformity norms and the uniform almost periodicity norms.

3.1 Gowers uniformity norms

The purpose of the Gowers uniformity norms is to quantify the notion of randomness in functions on $\mathbb{Z}_N$. Functions which have *small* Gowers uniformity norms will exhibit a lack of periodic behavior behavior. The precise definition is as follows.

Definition 3.1 (Gowers uniformity norms). *Let $f : \mathbb{Z}_N \to \mathbb{C}$ be a function. The dth Gowers uniformity norm $||f||_{U^d}$ is defined recursively for integers $d \geq 0$ by*

$$||f||_{U^0} := \mathbb{E}_{\mathbb{Z}_N} f$$

and

$$||f||_{U^d} := \left(\mathbb{E}_{h\in\mathbb{Z}_N}||\overline{f}T^h f||_{U^{d-1}}^{2^{d-1}}\right)^{1/2^d}.$$

The 0th and 1st Gowers uniformity norms are not true norms. However, for most of the arguments we will be taking $d = k-1$. Since we have already dealt with the simple cases of $k = 1$ and $k = 2$, we will not have to worry about this issue in most of the arguments to follow.

There are two useful interpretations suggested by Tao to understand the Gowers uniformity norms. As a heuristic, we can think of all of the Gowers uniformity norms as behaving something like the modulus of the average value of f over $\mathbb{Z}_N$. Indeed, the following equality looks quite similar similar to the preceding definition.

$$|\int_{\mathbb{Z}_N} f|^2 = \mathbb{E}_{h\in\mathbb{Z}_N}\int_{\mathbb{Z}_N}\overline{f}T^h f$$

Written in another way, $(\int_{\mathbb{Z}_N}\overline{f})\times(\int_{\mathbb{Z}_N} f) = \mathbb{E}_{h\in\mathbb{Z}_N}\int_{x\in\mathbb{Z}_N} f(x)f(x+h)$ so certainly this equality holds true. The first Gowers uniformity norm of f is of course exactly given by $|\mathbb{E}_{\mathbb{Z}_N} f|$.

Another useful way of thinking about the Gowers uniformity norm comes from finite Fourier series. The Fourier series of a function $f : \mathbb{Z}_N \to \mathbb{C}$ is given by

$$f(x) = \sum_{n=1}^{N} a_n e^{2\pi i x/N}, \qquad a_n = \int_{y\in\mathbb{Z}_N} f(y)e^{-2\pi i y n/N}$$

(Note that in the finite setting, questions about the convergence of Fourier series disappear. For more information, see [11]). Tao observes that the second Gowers uniformity norm is precisely equal to the l^4 norm of the Fourier coefficients of f.

$$||f||_{U^2} = \left(\sum_{n=1}^{N}|a_n|^4\right)^{1/4}$$

In other words, a small second Gowers uniformity norm corresponds to the state of having to small harmonics.

We conclude our discussion of the Gower's uniformity norms with several examples.

Example 3.1. If $f(x) = c$ is a constant function, then $||f||_{U^d} = |c|$ for all $d \geq 1$. This follows easily by induction. For the base case, we have

$$||f||_{U^1} = \left(\int_{h\in\mathbb{Z}_N}||\overline{c}T^h c||_{U^0}\right)^{1/2} = \left(\int_{h\in\mathbb{Z}_N}|||c|^2||_{U^0}\right)^{1/2} = \left(\int_{h\in\mathbb{Z}_N}|c|^2\right)^{1/2} = |c|$$

For the inductive step, we have

$$||f||_{U^d} = \left(\int_{h\in\mathbb{Z}_N}||\overline{c}T^h c||_{U^{d-1}}^{2^{d-1}}\right)^{1/2^d} = \left(\int_{h\in\mathbb{Z}_N}|\overline{c}c|^{2^{d-1}}\right)^{1/2^d} = |c|,$$

using above the shift invariance of the constant c and the inductive hypothesis.

Example 3.2. A simple example of a *quasiperiodic function*, discussed in the next section, is given by

$$W(x) := e^{2\pi i P(x)/N}$$

where $P(x)$ is a monic polynomial. If P is of degree m, then we have

$$||W||_{U^d} = 1, \qquad d > m.$$

To prove this, we argue by induction on m. For the base case, if $m = 0$, then $P(x) = 1$ and $W(x) = e^{2\pi i/N}$ is constant. Thus, for $d > 0$, we have $||W(x)||_{U^d} = 1$, by the result in example 3.1. For the induction step, suppose $P(x)$ is a monic polynomial of order m. We have

$$(\overline{W}T^h W)(x) = e^{2\pi i(P(x)-P(x+h))/N}$$

Observe that $P(x) - P(x+h)$ is a polynomial of order $m-1$, since the highest order terms cancel. Then by the induction hypotheses,

$$||\overline{W}T^h W||_{U^{d-1}} = 1$$

for all $d > m$. From this, it follows that

$$||W||_{U^d} = \mathbb{E}_{\mathbb{Z}_N} ||\overline{W}T^h W||_{U^{d-1}}^{2^{d-1}} = \mathbb{E}_{\mathbb{Z}_N} 1 = 1$$

as claimed.

3.2 Uniform almost periodicity norms

The purpose of the uniform almost periodicity norms is to quantify the appropriate notion of structure for proving theorem 1.1. In particular, the UAP^{k-2} quantify how closely a function behaves like (6), a quasiperiodic function of order $k-2$. To motivate the definition, we make the following observation. The shifts of the function $F(x)$, described by (6), take the form

$$T^n F(x) = \frac{1}{J} \sum_{j=1}^{J} c_j e^{2\pi i P_j(x+n)/N} = \mathbb{E}_{1 \le j \le J}(c_{n,j} g_j), \qquad n \in \mathbb{Z}_N, \tag{14}$$

where $c_{n,j}(x) = e^{2\pi i(P_j(x+n)-P_j(x))/N}$ and $g_j(x) = e^{2\pi i P_j(x)/N}$. Since $P_j(x)$ is a polynomial of degree d, the polynomial $P_j(x+n) - P_j(x)$ is of degree $d-1$, which means that $c_{n,j}(x)$ is a quasiperiodic function of order $d-1$, with modulus 1.[8] In short, every shift $T^n F$ may be written as a linear combination of the functions g_j, quasiperiodic of order d, with the coefficients $c_{n,j}$ taken to be quasiperiodic functions of one degree less. The appropriate generalization of functions of form (14), which

[8]Compare with example 3.2

we shall define, is the space of UAP^d functions, equipped with norm $||\cdot||_{UAP^d}$. Each UAP^d function $F : \mathbb{Z}_N \to \mathbb{C}$ will have the special property that its shifts take the form

$$T^n F = \frac{M}{|H|} \sum_{h \in H} c_{n,h}(x) g_h(x) = M\mathbb{E}(c_{n,h} g_h), \qquad n \in \mathbb{Z}_N,$$

where $M > 0$, H is some finite, non-empty set, h is a random variable taking values in H and each $c_{n,h}$ is a UAP^{d-1} function with $||c_{n,h}||_{UAP^{d-1}} \leq 1$ and each g_h is a bounded function.

We are now ready to launch into the formal definition of the space of UAP^d functions. To do so, we need introduce the concept of a *Banach algebra*.

Definition 3.2 (Banach algebra)**.** A vector subspace A of functions $f : \mathbb{Z}_N \to \mathbb{C}$, equipped with norm $|| \cdot ||_A$, is a *Banach algebra* if the following three properties are satisfied.

1. $\overline{f} \in A$ for all $f \in A$ (closed under conjugation).

2. $fg \in A$ for all $f, g \in A$ (closed under pointwise product).

3. $||fg||_A \leq ||f||_A ||g||_A$ for all $f, g \in A$.

We moreover say that the A is a *shift invariant* Banach algebra if $||T^n f||_A = ||f||_A$ for all $f \in A$ and all n. We say that A is a *scale invariant* Banach algebra if $||f_\lambda||_A = ||f||_A$ for all $f \in A$ and $\lambda \in \mathbb{Z}_N - \{0\}$, where f_λ is defined by $f_\lambda(x) = f(\lambda x)$.

Definition 3.3 (The $UAP[A]$ Banach algebra)**.** Given a shift-invariant Banach algebra A, we define the Banach algebra $UAP[A]$ to be the set of all functions $F : \mathbb{Z}_N \to \mathbb{C}$ whose shifts take the form

$$T^n F(x) = \frac{M}{|H|} \sum_{h \in H} c_{n,h}(x) g_h(x) = M\mathbb{E}(c_{n,h} g_h), \qquad n \in \mathbb{Z}_N, \tag{15}$$

where $M \geq 0$, H is some finite, non-empty set, h is a random variable taking values in H, $c := \{c_{n,h}\}_{n \in \mathbb{Z}_N, h \in H}$ is a collection of functions in A satisfying $||c_{n,h}||_A \leq 1$, and $g := \{g_h\}_{h \in H}$ is a collection of bounded functions. The norm $||F||_{UAP[A]}$ is defined to the infimum of all such M for which there exists a representation of form (15).

It is important to note that a space satisfying the properties given by definition 3.3 really is a Banach space. Moreover we can "compose" UAP Banach algebras to form new higher order UAP Banach algebras,

$$UAP^d[A] := \underbrace{UAP[UAP[...UAP[A]...]]}_{\text{d times}}.$$

Tao justifies these facts by proving the following proposition.

Proposition 3.1. *Let A be a shift invariant Banach algebra. The space $UAP[A]$ has the following properties.*

1. *The space $UAP[A]$ is also a shift invariant Banach algebra.*
2. *The space A is contained in $UAP[A]$.*
3. *For all $f \in A$, we have $||f||_A \geq ||f||_{UAP[A]}$.*
4. *If A is a scale invariant Banach algebra, then so is $UAP[A]$.*

This finally leads us to the definition of the space of uniform almost periodic functions of order d.

Definition 3.4 (UAP^d Banach algebra)**.** Let K denote the Banach algebra of all constant functions $c : \mathbb{Z}_N \to \mathbb{C}$, with norm defined to be the modulus $|c|$. The UAP^d Banach space is defined recursively by setting $UAP^0 := K$ and setting $UAP^d := UAP[UAP^{d-1}]$.

We conclude this section by noting that it is not very restrictive to require that functions be contained in a UAP^d Banach algebra. In fact, every function $f : \mathbb{Z}_N \to \mathbb{C}$ is contained in UAP^1 and, by proposition 3.1, the same function is contained in every higher order Banach algebra UAP^d, with $d \geq 1$. In particular, like the Gowers uniformity norms, the space of UAP^d functions with its associated norm has an interpretation which comes from finite Fourier series. Tao proves the following proposition.

Proposition 3.2. *Let $F : \mathbb{Z}_N \to \mathbb{C}$ be a function. Then $F \in UAP^1$, and its norm satisfies*

$$||F||_{UAP^d} = \sum_{n=1}^{N} |a_n|, \quad a_n = \int_{y \in \mathbb{Z}_N} F(y) e^{-2\pi i y n/N}.$$

In other words, the UAP^1 norm is precisely equal to the l^1 norm of the Fourier coefficients of F.

3.3 The relationship between Gowers uniform and uniform almost periodic functions

We are now ready to explain our claim that Gowers uniform functions are, in a sense, orthogonal to uniform almost periodic functions. Recall that the inner product of two functions in $L^2(\mathbb{Z}_N)$ is defined by $\langle g, h \rangle := \int_{\mathbb{Z}_N} \overline{g} h$. If f is a Gowers uniform function and F is a uniform almost periodic function, one cannot say in general that $\langle f, F \rangle = 0$, since the term "Gowers uniform" is fuzzy, just indicating that the Gowers uniformity norm of a function is "small". However, Tao does prove the following proposition, which is almost as good.

Proposition 3.3 (Uniformity is orthogonal to almost periodicity). *Given any functions f, F with $F \in UAP^{k-2}$ for some $k \geq 2$, we have*

$$|\langle f, F \rangle| \leq ||f||_{U^{k-1}} ||F||_{UAP^{k-2}}.$$

In other words, whenever $||f||_{U^{k-1}}$ is small, $\langle f, F \rangle$ is close to zero, which is to say that f is "almost" orthogonal to F.
The next result is a partial converse to proposition 3.2. We state it as a lemma, because it plays a needed role in Tao's proof of Szemerédi's theorem.

Lemma 3.1 (Lack of Gowers Uniformity implies correlation with a UAP function). *For a bounded function f and some $k \geq 3$ and $\epsilon > 0$, suppose that $||f||_{U^{k-1}} \geq \epsilon$. Then there exists a bounded function $F \in UAP^{k-2}$ with $||F||_{UAP^{k-2}} \leq 1$ such that $|\langle f, F \rangle| \geq \epsilon^{2^{k-1}}$.*

This says intuitively that if a function lacks Gowers uniformity (of degree $k-1$), even by a small amount, then it is not completely orthogonal to the space of UAP^{k-2} functions.

4 An alternative proof of the generalized Von Neumann theorem using UAP norms

To prove theorem 2.2, Tao gives an inductive argument which has the advantage of being simple and elegant, relying on little more than the Cauchy-Schwarz inequality. However, Tao also mentions the possibility of an alternative proof to theorem 2.2, which uses UAP^d norms and the orthogonality condition established by proposition 3.3. We work out the alternative proof in this section. It is worthwhile seeing the alternative argument because it allows us to frame theorem 2.2 within the randomness-structure schema of the rest of the paper. It will also give us a better handle on how the Gowers uniformity and uniform almost periodicity norms are used in practice.

For the purpose of proving theorem 2.2, we can make several simplifying assumptions about (7). First, by reindexing, can assume that the function in $\{f_0, ..., f_{k-1}\}$ with the minimal Gowers uniformity norm is f_0. Second, the map $T^{-\lambda_0 r}$ preserves expectation, by proposition 2.1. Thus we can assume that $\lambda_0 = 0$. Under these assumptions, we have

$$\mathbb{E}_{r \in \mathbb{Z}_N} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{\lambda_j r} f_j = \int_{\mathbb{Z}_N} f_0 \mathbb{E}_{r \in \mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j = \left\langle \overline{f_0}, \mathbb{E}_{r \in \mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j \right\rangle \quad (16)$$

Theorem 2.2 becomes a quick consequence of the following lemma.

Lemma 4.1. *If $k \geq 2$ and $\lambda_1, ..., \lambda_{k-1}$ are distinct elements in $\mathbb{Z}_N$, then for any bounded functions $f_1, ..., f_{k-1} : \mathbb{Z}_N \to \mathbb{C}$, we have*

$$\mathbb{E}_{r \in \mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j \in UAP^{k-2}. \tag{17}$$

with a UAP^{k-2} norm less than 1.

Indeed, once this lemma has been established, we need only apply proposition 3.3 to (16), along with the fact that the Gowers uniformity norms are invariant under conjugation, to obtain the desired result

$$\left| \mathbb{E}_{r \in \mathbb{Z}_N} \int_{\mathbb{Z}_N} \prod_{j=0}^{k-1} T^{\lambda_j r} f_j \right| \leq ||\overline{f_0}||_{U^{k-1}} \left|\left| \mathbb{E}_{r \in \mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j \right|\right|_{UAP^{k-2}} \leq ||f_0||_{U^{k-1}}. \tag{18}$$

It remains to prove lemma 5.1.

Proof of lemma 5.1. By reindexing we can assume that $\lambda_1 \neq 0$. Another simplifying assumption is made possible by proposition 2.2. This proposition tells us that the map $r \mapsto \lambda_1 r (\text{mod} N)$ takes $\mathbb{Z}_N$ bijectively to itself. (Recall that addition is taken modulo N, and N is assumed to be a large prime number.) Thus by reindexing the f_j's to match the natural ordering of $0, \lambda_1, 2\lambda_1, ..., (N-1)\lambda_1$ we can assume that $\lambda_1 = 1$.

We proceed by induction on k. For the base case, take $k = 2$. In this case, we have

$$(17) = \mathbb{E}_{r \in \mathbb{Z}_N} T^r f_1(x) = \frac{1}{N}(f_1(x) + f_1(x+1) + \cdots + f_1(x + (N-1))). \tag{19}$$

This quantity does not depend on integer x and thus is constant. This means that (17) is contained in UAP^0 when $k = 2$. Moreover, by boundedness of f_1, we have $|\mathbb{E}_{r \in \mathbb{Z}_N} T^{\lambda_1 r} f_1(x)| \leq 1$, and this means that the norm of (17) when $k = 2$ is at most 1.

Now we handle the induction step. Suppose (17) holds true for $k - 1$. We compute

$$\mathbb{E}_{r \in \mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j = \mathbb{E}_{r \in \mathbb{Z}_N} \left(T^r f_1 \prod_{j=2}^{k-1} T^{\lambda_j r} f_j \right),$$

$$= \mathbb{E}_{r \in \mathbb{Z}_N} T^s \left(T^r f_1 \prod_{j=2}^{k-1} T^{\lambda_j r} f_j \right) = \frac{1}{N} \sum_{r=0}^{N-1} \left(T^{r+s} f_1 \prod_{j=2}^{k-1} T^{\lambda_j r + s} f_j \right)$$

$$= \frac{1}{N} \sum_{r=0}^{N-1} \left(T^r f_1 \prod_{j=2}^{k-1} T^{\lambda_j (r-s)+s} f_j \right) = \mathbb{E}_{r \in \mathbb{Z}_N} \left(T^r f_1 \prod_{j=2}^{k-1} T^{\lambda_j (r-s)+s} f_j \right)$$

$$= \mathbb{E}_{s\in\mathbb{Z}_N}\mathbb{E}_{r\in\mathbb{Z}_N}\left(T^r f_1 \prod_{j=2}^{k-1} T^{\lambda_j(r-s)+s} f_j\right)$$

$$= \mathbb{E}_{r\in\mathbb{Z}_N}\left((T^r f_1)\mathbb{E}_{s\in\mathbb{Z}_N} \prod_{j=2}^{k-1} T^{\lambda_j(r-s)+s} f_j\right) \tag{20}$$

In the first line above, we use the fact that fact that $\lambda_1 = 1$ to pull out $T^r f_1$. In the second line, we introduce the shift T^s, for an $s \in \mathbb{Z}_N$, which leaves the expected value unchanged, by proposition 2.1. In the third line we make the change of indices $r \mapsto r - s$.[9] In the fourth line we take the average over $s \in \mathbb{Z}_N$. In the fifth line we pull this average inside, noting that $T^r f_1$ does not depend on s.

Applying the shift map T^n to (20) thus gives us

$$T^n \mathbb{E}_{r\in\mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j = \mathbb{E}_{r\in\mathbb{Z}_N}\left((T^{r+n} f_1)\mathbb{E}_{s\in\mathbb{Z}_N} \prod_{j=2}^{k-1} T^{\lambda_1(r-s)+s+n} f_j\right)$$

$$= \mathbb{E}_{r\in\mathbb{Z}_N}\left((T^r f_1)\mathbb{E}_{s\in\mathbb{Z}_N} \prod_{j=2}^{k-1} T^{\lambda_1(r-n-s)+s+n} f_j\right)$$

$$= \mathbb{E}_{r\in\mathbb{Z}_N}\left((T^r f_1)\mathbb{E}_{s\in\mathbb{Z}_N} \prod_{j=2}^{k-1} T^{(1-\lambda_1)s} f_j\right), \tag{21}$$

here making the change of variables $r \mapsto r - n$ and then applying the shift map $T^{-\lambda_1(r-n)+n}$ inside $\mathbb{E}_{s\in\mathbb{Z}_N}(\cdot)$, which leaves the expectation invariant. We claim that (21) has the form (15) with $M = 1$. Indeed, $T^r f_1$ is a bounded, complex valued function on $\mathbb{Z}_N$, and $\mathbb{E}_{s\in\mathbb{Z}_N} \prod_{j=2}^{k-1} T^{(1-\lambda_1)s} f_j \in UAP^{K-3}$ by the induction assumption. Thus $T^n \mathbb{E}_{r\in\mathbb{Z}_N} \prod_{j=1}^{k-1} T^{\lambda_j r} f_j$ is contained in UAP^{k-2} with UAP^{k-2} norm less than or equal to 1. This completes the induction step, so we are done. □

5 Energy incrementation

We present Tao's energy incrementation argument for proof of the structure theorem (theorem 2.3).

5.1 σ-algebras and the $L^2(\mathcal{B})$ Hilbert space structure

A central idea behind Tao's proof of the structure theorem is to not just work with uniform almost periodic functions, but also with σ-algebras which they generate.

[9]It is extremely important to note that we are working in the cyclic group $\mathbb{Z}_N$ to make sense of these index changes. Observe that the sum still runs from 0 to $N - 1$.

σ-algebras are sets of subsets of $\mathbb{Z}_N$ which play a role analogous to partitions of sets of real numbers in the theory of Riemann integration. The goal in this section is to present two propositions which allow us to approximate a uniform almost periodic function by its mean value on the minimal elements of a σ-algebra. We begin with some definitions.

Definition 5.1 (σ-algebra, atom)**.** A set $\mathcal{B}$ of subsets of $\mathbb{Z}_N$ is a called a σ-algebra of $\mathbb{Z}_N$ if the following conditions are satisfied.

1. $\emptyset \in \mathcal{B}$ and $\mathbb{Z}_N \in \mathcal{B}$.
2. If $A, B \in \mathcal{B}$, then $A \cap B \in \mathcal{B}$ and $A \cup B \in \mathcal{B}$ and $A - B \in \mathcal{B}$ (in words, $\mathcal{B}$ is closed under unions, intersections, and complementation).

A set $A \in \mathcal{B}$ is said to be an atom of $\mathcal{B}$ if A has no subsets also contained in $\mathcal{B}$.

Note that the atoms of $\mathcal{B}$ must form a partition of $\mathbb{Z}_N$.

In analogy with the union operation in set theory or the direct sum of two vector spaces in linear algebra, we would like to have a way of combining two σ-algebras to produce another σ-algebra. This is achieved by the following operation.

Definition 5.2 ($\vee$-sum of σ-algebras)**.** *Given two σ algebras $\mathcal{B}_1$ and $\mathcal{B}_2$ we denote by $\mathcal{B}_1 \vee \mathcal{B}_2$ the smallest σ-algebra which contains the elements of both $\mathcal{B}_1$ and $\mathcal{B}_2$.*

It is important to note that the "smallest" σ-algebra is here uniquely determined, and so the operator $\vee$ is well-defined. In particular, if $\mathcal{B}$ is a minimal σ-algebra then it is precisely equal to the set $\{B \cup C, B \cap C, B - C : B, C \in \mathcal{B}_1 \cup \mathcal{B}_2\}$.

Just as functions on the real numbers may be Riemann measurable depending on (roughly speaking) how well the function may be approximated on partitions of its domain, we also have the following concept of measurability with respect a σ-algebra for functions $\mathbb{Z}_N \to \mathbb{C}$. With this concept of measurability, we can define the Hilbert space $L^2(\mathcal{B})$ of functions measurable with respect to a σ-algebra $\mathcal{B}$. This will be helpful for interpreting geometrically some of the ideas to follow.

Definition 5.3 (Measurability, the $L^2(\mathcal{B})$ Hilbert space)**.** A function f is said to be *measurable* with respect to a σ-algebra $\mathcal{B}$ if all the level sets of f are contained in $\mathcal{B}$. We define $L^2(\mathcal{B})$ to be the subspace of $L^2(\mathbb{Z}_N)$ consisting of all functions which are measurable with respect to the σ-algebra $L^2(\mathcal{B})$.

It is not difficult to verify that $L^2(\mathcal{B})$ is a Hilbert subspace, and we won't prove it here. We next define the expectation operator $f \to \mathbb{E}(f|\mathcal{B})$. The expectation $\mathbb{E}(f|\mathcal{B})$, representing the best approximation of f by functions measurable in $\mathcal{B}$, is defined to be the orthogonal projection of f onto the space $L^2(\mathcal{B})$,

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

where $\mathcal{B}(x)$ is the (necessarily unique) atom of $\mathcal{B}$ which contains x. One simple fact, which will be useful later in our argument, is the following.

$$\int_{\mathbb{Z}_N} \mathbb{E}(f|\mathcal{B}) = \int_{\mathbb{Z}_N} f \tag{22}$$

This follows immediately by expanding $\mathbb{E}(f|\mathcal{B}) = \mathbb{E}_{\mathcal{B}(x)} f$ as the sum definition for expected value. To verify the fact that $\mathbb{E}(f|\mathcal{B})$ is an orthogonal projection, recall that for functions g, h in the Hilbert space $L^2(\mathbb{Z}_N)$ we had the inner product $\langle g, h\rangle := \int_{\mathbb{Z}_N} \bar{g}h$. We have

$$\langle \mathbb{E}(f|\mathcal{B}), f - \mathbb{E}(f|\mathcal{B})\rangle = \int_{\mathbb{Z}_N} f\mathbb{E}(\overline{f}|\mathcal{B}) - \int_{\mathbb{Z}_N} \mathbb{E}(f|\mathcal{B})\mathbb{E}(\overline{f}|\mathcal{B})$$

$$= \int_{\mathbb{Z}_N} \mathbb{E}(f\mathbb{E}(\overline{f}|\mathcal{B})|\mathcal{B}) - \int_{\mathbb{Z}_N} \mathbb{E}(f|\mathcal{B})\mathbb{E}(\overline{f}|\mathcal{B}) = 0.$$

Building on this fact, we will say that a function f is *orthogonal* to a σ-algebra $\mathcal{B}$ if $\mathbb{E}(f|\mathcal{B}) = 0$. Thus any function $f \in L^2(\mathbb{Z}_N)$ has the decomposition,

$$f = (f - \mathbb{E}(f|\mathcal{B})) + \mathbb{E}(f|\mathcal{B}) \tag{23}$$

where $(f - \mathbb{E}(f|\mathcal{B}))$ is orthogonal to $\mathcal{B}$ and $\mathbb{E}(f|\mathcal{B})$ is measurable in $\mathcal{B}$.

Uniform almost periodic functions are related to σ-algebras by the following proposition. It says that the level sets of any function in UAP^d may be arbitrarily well approximated by the atoms of some σ-algebra and that functions measurable in this σ-algebra may be arbitrarily well approximated by functions in UAP^d.

Proposition 5.1. *(UAP functions generate compact σ-algebras) Fix $d \geq 0$, and suppose $G \in UAP^d$ with the bound $||G||_{UAP^d} \leq M$ for some $M > 0$. Then for any $\epsilon > 0$ we can find a σ-algebra $\mathcal{B}_\epsilon(G)$ with $O_{M,\epsilon}(1)$ atoms, which satisfies*

$$||G - \mathbb{E}(G|\mathcal{B}_\epsilon(G) \vee \mathcal{B})||_{L^\infty} = O(\epsilon)$$

for any σ-algebra $\mathcal{B}$. Moreover, for any bounded, nonnegative function $f \in L^2(\mathcal{B}_\epsilon(G))$ and any $\delta > 0$, there is a bounded, nonnegative function $f_{UAP} \in UAP^d$ such that

$$||f - f_{UAP}||_{L^2} \leq \delta$$

and

$$||f_{UAP}||_{UAP^d} = O_{M,\epsilon,\delta}(1).$$

To describe the energy incrementation argument in the next subsection, the concept of the *complexity* of a σ-algebra is crucial. Roughly speaking, the complexity of a σ-algebra $\mathcal{B}$ represents minimal amount of work required to build up $\mathcal{B}$ from σ-algebras generated by uniform almost periodic functions of order d. We imagine that we would like to be as lazy as possible, building up $\mathcal{B}$ from just a few σ-algebras, each with just a few atoms and generated by small UAP^d functions. With more complex σ-algebras, it is more difficult to accomplish this task.

Definition 5.4 (Compact σ-algebra, complexity)**.** A σ-algebra $\mathcal{B}$ is said to be compact of *order* d with *complexity at most* X if

$$\mathcal{B} = \mathcal{B}_{\epsilon_1}(G_1) \vee \cdots \vee \mathcal{B}_{\epsilon_K}(G_K)$$

where $G_j \in UAP^d$, and $||G_j||_{UAP^d} \leq X$, and $\epsilon_j \leq 1/(1+X)$, and $1 \leq K \leq X$ $(j = 1, ..., K)$. The *complexity* of $\mathcal{B}$ is defined to be the minimal X for which the above conditions hold true.

The next proposition describes how well a function in a given σ-algebra may be approximated by functions in UAP^d.

Proposition 5.2. *Fix $d \geq 0$ and $X \geq 0$, and suppose that $\mathcal{B}$ is a σ-algebra compact of order d with complexity at most X, then for any bounded, nonnegative function $f \in L^2(\mathcal{B})$ and any $\delta > 0$, there is a bounded, nonnegative function $f_{UAP} \in UAP^d$ such that*

$$||f - f_{UAP}||_{L^2} \leq \delta$$

and

$$||f_{UAP}||_{UAP^d} = O_{d,\delta,X}(1).$$

5.2 Abstract energy incrementation argument

Important to Tao's proof of the structure theorem is the concept of the energy of a σ-algebra, which measures the size of the best approximation of a set of functions on a given σ-algebra.

Definition 5.5 (Energy)**.** The *energy* $\mathcal{E}_f(\mathcal{B})$ of an m-tuple of functions $f = (f_1, ..., f_m)$ on a σ-algebra $\mathcal{B}$ is defined by

$$\mathcal{E}_f(\mathcal{B}) = \sum_{j=0}^{m} ||\mathbb{E}(f_j|\mathcal{B})||_{L^2}^2 \tag{24}$$

Note that if $\mathcal{B}' \supset \mathcal{B}$ is a finer σ-algebra than $\mathcal{B}$, then $\mathbb{E}(f_j|\mathcal{B}')$ is orthogonal to $\mathbb{E}(f_j|\mathcal{B}) - \mathbb{E}(f_j|\mathcal{B}')$. The Pythagorean theorem then gives us the useful formula

$$\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B}) = \sum_{j=0}^{m} ||\mathbb{E}(f_j|\mathcal{B}') - \mathbb{E}(f_j|\mathcal{B})||_{L^2}^2 \tag{25}$$

As an immediate consequence of this formula, we note that the energy of a σ-algebra always increases with the fineness of the σ-algebra.

The abstract energy incrementation lemma reduces the matter proving some property $P(M)$, for some $M > 0$, to a matter proving a dichotomy – either $P(M)$ is true relative to some σ-algebra $\mathcal{B}$, or we can find a finer σ-algebra $\mathcal{B}'$ which satisfies certain properties.

Lemma 5.1 (Abstract energy incrementation)**.** *Suppose we have a property $P(M)$ which depends on parameter $M > 0$. Fix $d \geq 0$, and let $f = (f_1, ..., f_m)$ be a collection of bounded functions.*

Given any $X, X' > 0$, and given any σ-algebras $\mathcal{B}$, compact of order d with complexity at most X, and $\mathcal{B}'$ finer than $\mathcal{B}$, also compact of order d with complexity at most X', suppose at least one of the following conditions is true:

i. If we have the energy gap condition $\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B}) \leq \tau^2$, then $P(M)$ is true for some $M = O_{m,\tau,X,X'}(1)$.

ii. There exists a σ-algebra finer than $\mathcal{B}$ and compact of order d with complexity at most $O_{d,\tau,X,X'}$, which satisfies the energy increment property $\mathcal{E}_f(\mathcal{B}'') - \mathcal{E}_f(\mathcal{B}') \geq c(d, \tau, X)$, where positive quantity $c(d, \tau, X) > 0$.

Then $P(M)$ is true for some $M = O_{m,d,\tau}(1)$.

Proof. Consider the following algorithm.

1. Set σ-algebra $\mathcal{B} = \{0, \mathbb{Z}_N\}$ and let X denote the complexity of $\mathcal{B}$.
2. Set σ-algebra $\mathcal{B}' = \mathcal{B}$ and let X' denote the complexity of $\mathcal{B}'$.
3. Since $\mathcal{B}$ and $\mathcal{B}'$ satisfy the energy gap condition in (i), by assumption either $P(M)$ is true for some $M = O_{m,d,X,X'}(1)$, or we can find a finer σ-algebra $\mathcal{B}''$ with complexity at most $O_{m,d,X,X'}(1)$ satisfying condition (ii). If the former is true, stop. If the latter is true, go on to step 4.
4. If $\mathcal{B}$ and $\mathcal{B}'$ satisfy the energy gap condition $\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B}) \leq \tau^2$, replace $\mathcal{B}'$ with $\mathcal{B}''$ and go to step 3 again. Otherwise, replace $\mathcal{B}$ with $\mathcal{B}''$ and go to step 2 again.

It follows from our assumptions that each time we replace $\mathcal{B}'$ with $\mathcal{B}''$, $\mathcal{E}_f(\mathcal{B}')$ increases by at least $c(d, \tau, X)$. As soon as $\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B})$ exceeds τ^2, the algorithm requires $\mathcal{B}$ be replaced by $\mathcal{B}''$. Thus the algorithm takes at most $\tau^2/c(d, \tau, X) = O_{d,\tau,X}(1)$ steps before replacing $\mathcal{B}$ with $\mathcal{B}''$. Fundamental to this argument, the complexity of the resulting $\mathcal{B}''$ depends on X, and not on any of the intermediate complexities X'; indeed, this part of the algorithm starts out with $\mathcal{B}' = \mathcal{B}$ (and hence $X' = X$) and keeps replacing $\mathcal{B}'$ with another σ-algebra which has complexity $O_{m,d,X,X'}(1)$, so when the energy gap violates $\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B}) \leq \tau^2$, the resulting $\mathcal{B}''$ has complexity depending only on d,τ,X, and the number of steps for the energy gap condition to be violated. Since we have seen that the number of steps is $O_{d,\tau,X}(1)$, we conclude that the complexity of the resulting $\mathcal{B}''$ is $O_{d,\tau,X}(1)$.

Now consider the part of the algorithm where we replace $\mathcal{B}$ by $\mathcal{B}''$. By boundedness of the f_j's for $j = 1, ..., m$, the energy $\mathcal{E}_f(\mathcal{B})$ will never exceed m. Since each time $\mathcal{B}$ is replaced by $\mathcal{B}''$ only when the energy gap increases by more than τ^2, the replacement can occur only $m/\tau^2 = O_{m,\tau}(1)$ times.

Combining these results we conclude that the algorithm iterates at most $O_{d,m,\tau,X}(1)$ times before stopping. Since each time $\mathcal{B}''$ replaces $\mathcal{B}$, the complexity increases by at most $O_{d,\tau,X}(1)$, the complexity X is determined by the initial complexity of $\{\emptyset, \mathbb{Z}_N\}$ which is zero, so in fact the complexity increases by $O_{m,\tau}(1)$. In particular, the M at which the algorithm terminates is at most $O_{m,d,\tau}(1)$, which is what we needed to show. □

5.3 Energy incrementation applied to the structure theorem

By the previous lemma, it is sufficient to prove the following dichotomy to show that the structure theorem is true.

Lemma 5.2 (Structure theorem dichotomy)**.** *Let $k \geq 0$, and suppose we have the σ-algebras $\mathcal{B}$ and $\mathcal{B}'$ both of order $k-2$, with complexities of at most X and X' respectively. Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be a function. Let f be a nonnegative bounded function which, for some $\delta > 0$, satisfies $\int_{\mathbb{Z}_N} f \geq \delta$ and the energy gap condition,*

$$\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B}) \leq \tau^2.$$

for $\tau = \delta^2/5000k$. Then at least one of the following is true

i. We can find an $M = O_{k,\delta,X}(1)$ such that the structure theorem (2.3) is true for the f and δ given above and for some functions $f_{U\perp}$, f_{UAP}, and $d := k-2$ (here preserving all the notation in the statement of theorem 3.5).

ii. There exists a σ-algebra $\mathcal{B}''$ finer than $\mathcal{B}$ and compact of order d with complexity at most $O_{d,\tau,X,X'}$, which satisfies the energy increment property $\mathcal{E}_f(\mathcal{B}'') - \mathcal{E}_f(\mathcal{B}') \geq c(d,\tau,X)$, where positive quantity $c(d,\tau,X,F) > 0$ independent of X.

Proof. Start by fixing σ-algebras $\mathcal{B} \subset \mathcal{B}'$ in $\mathbb{Z}_N$, both compact of order $k-2$ with complexity at most $O(X)$. We split f into two orthogonal parts

$$f = f_U + f_{U^\perp}$$

where $f_{U^\perp} = \mathbb{E}(f|\mathcal{B}')$ and $f_U = f - \mathbb{E}(f|\mathcal{B}')$. Since f is non-negative and bounded, $\mathbb{E}(f|\mathcal{B})$ is also non-negative and bounded. So by proposition 4.2, there is a function $f_{UAP} \in UAP^{k-2}$ such that

$$||\mathbb{E}(f|\mathcal{B}) - f_{UAP}||_{L^2} \leq \frac{\delta^2}{5000k} \tag{26}$$

and

$$||f_{UAP}||_{UAP^d} < M$$

for some $M = O_{k,\delta,X}(1)$. We can show that functions $f_{U^\perp}, f_{UAP}$ satisfy the estimates (8),(9), and (10) for $d = k-2$. Indeed, with $\tau = \delta^2/5000k$, we have by assumption the energy gap condition $\mathcal{E}_f(\mathcal{B}') - \mathcal{E}_f(\mathcal{B}) \leq \tau^2$, and thus by (25) $||E(f|\mathcal{B}') - E(f|\mathcal{B})||_{L^2} \leq \tau$. Combined with (26) and the triangle inequality, this gives us $||\mathbb{E}(f|\mathcal{B}) - f_{UAP}||_{L^2} \leq 2\tau$, or

$$||f_{U^\perp} - f_{UAP}||_{L^2} \leq 2\tau < \frac{\delta^2}{1024k}$$

So (8) is satisfied. To see that (9) is also satisfied, note that by (22) $\int_{\mathbb{Z}_N} f_{U^\perp} = \int_{\mathbb{Z}_N} \mathbb{E}(f|\mathcal{B}') = \int_{\mathbb{Z}_N} f \geq \delta$, by hypothesis. We just showed that (10) is satisfied. If (11) is satisfied, then the first half (i) of the structure theorem dichotomy is true, so let us assume that (11) is false, and show that this implies (ii) is true. This assumption gives us

$$||f_U||_{U^{k-1}} > \frac{1}{F(M)}$$

But then Lemma 3.1 implies that there is a function $G \in UAP^{k-2}$ such that $||G||_{U^{k-2}} \leq 1$ and $|\langle f_U, G\rangle| \geq c_0(k, \delta, M, F) > 0$. Let $\mathcal{B}'' = \mathcal{B}' \vee \mathcal{B}_\epsilon(G)$, for a given $\epsilon = \epsilon(k, \delta, M)$ to be chosen later. We make yet another split into orthogonal parts

$$f_U = (f - \mathbb{E}(f|\mathcal{B}'')) + (\mathbb{E}(f|\mathcal{B}'') - \mathbb{E}(f|\mathcal{B}'))$$

$$G = (G - \mathbb{E}(G|\mathcal{B}'')) + (\mathbb{E}(G|\mathcal{B}'') - \mathbb{E}(G|\mathcal{B}')) + \mathbb{E}(G|\mathcal{B}')$$

Observe that (by (23)) the first of the orthogonal parts for f_U is orthogonal to $\mathcal{B}''$ (and thus also to the courser σ-algebra $\mathcal{B}'$), while the second term is measurable in $\mathcal{B}''$. Similarly the first of the orthogonal parts for G is orthogonal to $\mathcal{B}''$ (and to $\mathcal{B}'$), while the second part is measurable in $\mathcal{B}''$ and the third part is measurable in $\mathcal{B}'$. This causes the inner product of f_U and G to a take the simple form

$$\langle f_U, G\rangle = \langle f - \mathbb{E}(f|\mathcal{B}'')), G - \mathbb{E}(G|\mathcal{B}'')\rangle + \langle \mathbb{E}(f|\mathcal{B}'') - \mathbb{E}(f|\mathcal{B}'), \mathbb{E}(G|\mathcal{B}'') - \mathbb{E}(G|\mathcal{B}')\rangle$$

By proposition 5.1, $||G - \mathbb{E}(G|\mathcal{B}' \vee \mathcal{B}_\epsilon(G))||_{L^\infty} = O(\epsilon)$. Since f is bounded, we therefore have

$$|\langle f - \mathbb{E}(f|\mathcal{B}'')), G - \mathbb{E}(G|\mathcal{B}'')\rangle| = O(\epsilon).$$

Considering the lower bound on the inner product $\langle f_U, G\rangle$, we conclude that for a sufficiently small ϵ,

$$||\mathbb{E}(f|\mathcal{B}'') - \mathbb{E}(f|\mathcal{B}'')||_{L^2} \geq c_0(k, \delta, M, F),$$

which combined with (25) gives us the energy increment and implies (ii) is true. □

6 Conclusion

Number theory remains ever old and ever new, a fact testified by the variety of interesting proofs of Szemeredi's theorem over the past forty years. The quantitative ergodic theory proof of Szemerédi's theorem given by Tao [14] provides us with fascinating, if somewhat unexpected, way to understand the theorem, and it is full of ideas which are interesting in their own right. In particular, the Gowers uniformity norms paired with the uniform almost periodicity norms provide a useful perspective for studying systems which exhibit both structured and quasi-random behavior. Given that similar methods have been used to solve other difficult problems in recent years, for example Green and Tao in [7], understanding the randomness-structure dichotomy holds great promise for the future.

References

[1] Barreira, Luis and Valls, Claudia. Dynamical Systems: an introduction. Springer, 2012.

[2] Einseidler, Manfred and Ward, Thomas. Ergodic Theory with a view towards Number Theory. Springer, 2011.

[3] Erdös, Paul and Turán, Paul. On some sequences of integers. Journal of the London Mathematical Society 11(4) 261-264, 1936.

[4] Furstenberg, Hillel. Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. Journal d'Analyse Mathématique 31(1977) 204-256.

[5] Gowers, Timothy. A new proof of Szemerédi's theorem for arithmetic progressions of length four. GAFA, Geometric and Functional Analysis 8(1998) 528-551.

[6] Gowers, Timothy. A new proof of Szemerédi's theorem. GAFA, Geometric and Functional Analysis 11(2001) 465-588.

[7] Green, Ben and Tao, Terence. The primes contain arbitrarily long arithmetic progressions. Annals of Mathematics 167(2) 481-547, 1 March 2008.

[8] Judson, Thomas. Abstract Algebra: Theory and Applications. Self-published, 2013.

[9] Kra, Bryna. Ergodic methods in additive combinatorics. Centre de Recherche Mathématique, CRM Proceedings and Lecture Notes, 2006. Web http://www.math.-northwestern.edu/ kra/papers/montreal.pdf.

[10] Roth, Klaus. On certain sets of integers. Journal of the London Mathematical Society 28(1953) 245-252.

[11] Stein, Elias M. and Shakarchi, Rami. Fourier Analysis: An Introduction. Princeton Lectures in Analysis, Bk.III. Princeton: Princeton University Press, 2003.

[12] Stein, Elias M. and Shakarchi, Rami. Real Analysis: Measure Theory, Integration, and Hilbert Spaces. Princeton Lectures in Analysis, Bk.III. Princeton: Princeton University Press, 2005.

[13] Szemerédi, Endré. On sets of integers containing no k elements in arithmetic progression. Acta Arithmetica 27(1975) 199-245.

[14] Tao, Terence. A quantitative ergodic theory proof of Szemerédi's theorem. The Electronic Journal of Combinatorics 13 (2006), #R99.

[15] Varnavides, P. On certain sets of positive density. Journal of the London Mathematical Society 39(1959) 358-360.