TECHNIQUES IN COMBINATORICS – LECTURE NOTES

W. T. GOWERS

1. Introduction

The aim of this course is to equip you with a set of tools that will help you solve certain combinatorial problems much more easily than you would be able to if you did not have these tools. So, as the title of the course suggests, the emphasis will be much more on the methods used to prove theorems than on the theorems themselves: my main reason for showing you proofs of theorems is to draw attention to and illustrate the methods.

I do not normally write lecture notes, but am prompted to do so by the fact that some people at the first lecture were unable to take notes because there was not enough space to sit down. I have not decided whether I will write notes for the whole course, but I'll make what I write available soon after writing it. One effect of this will be that the the notes may well come to an abrupt end while I'm in the middle of discussing something.

2. The average lies between the minimum and the maximum

One of the simplest tools to describe – though it sometimes requires some ingenuity to use – is the principle that a random variable cannot always be less than its mean and cannot always be greater than its mean. More symbolically,

$$\inf X < \mathbb{E}X < \sup X.$$

By and large, we will be dealing with finite sample spaces, so the inf and sup are a min and a max.

Perhaps the most famous example of a proof that uses this principle is Erdős's lower bound for the Ramsey number $R(k,k)$.

Theorem 2.1. *Let $k, n \in \mathbb{N}$ be such that $k \geq 2$ and $n \leq 2^{(k-1)/2}$. Then there exists a graph with n vertices that contains no clique or independent set of size k.*

Proof. Let the edges of G be chosen randomly. That is, each edge is chosen with probability $1/2$ and all choices are independent.

For any k vertices $x_1, \dots, x_k$, the probability that they span a clique or independent set is $2.2^{-\binom{k}{2}}$. Therefore, the expected number of cliques or independent sets of size k is $2\binom{n}{k}2^{-\binom{k}{2}}$. (Note that we have just used linearity of expectation – the fact that this does not require the random variables in question to be independent is often extremely helpful.)

By the basic principle above, there must exist a choice of edges such that the number of cliques or independent sets of size k is at most $2\binom{n}{k}2^{-\binom{k}{2}}$, so if we can ensure that this is less than 1, then we will be done.

But $\binom{n}{k} < n^k/2$, so $2\binom{n}{k}2^{-\binom{k}{2}} < n^k 2^{-k(k-1)/2}$. Therefore, if $n \leq 2^{(k-1)/2}$ we are done, as claimed. □

The basic ideas of the proof above could be summarized as follows.

(1) Choose the graph randomly.
(2) Calculate the expected number of cliques/independent sets.
(3) Adjust n until the expected number falls below 1.

It is very important to learn to think of proofs in this sort of condensed way, trusting that you can do the necessary calculations.

Here is a second example – also a theorem of Erdős. Call a set of integers *sum free* if it contains no three elements a, b, c with $a + b = c$.

Theorem 2.2. *Let X be a set of n positive integers. Then X has a sum-free subset Y of size at least $n/3$.*

Before I give the proof, let me try to guess how Erdős thought of it. He might have begun by thinking about ways of ensuring that sets are sum free. Perhaps he started with the simple observation that the set of all odd numbers is sum free. Unfortunately, that doesn't help much because X might consist solely of even numbers. But we could try other moduli.

What would the basic strategy be? What we'd like to do is find a large collection of pretty dense sum-free sets that is in some sense evenly distributed, so that we can argue that X must have a large intersection with one of them, and therefore have a large sum-free subset. If we think about unions of residue classes mod p we soon spot that the interval $[p/3, 2p/3]$ is sum free, and having spotted that we realize that any non-zero multiple of that interval (mod p) will have the same property. That gives us a collection of sets of density $1/3$ and each non-multiple of p is contained in the same number of sets (which is what I meant by "in some sense evenly distributed"). That's basically the proof, but here are the details, expressed slightly differently.

Proof. Let $p > \max X$ and let a be chosen randomly from $\{1, 2, \dots, p-1\}$. Let $[p/3, 2p/3]$ denote the set of integers mod p that lie in the real interval $[p/3, 2p/3]$. Thus, if p is of the form $3m+1$, then $[p/3, 2p/3] = \{m+1, \dots, 2m\}$ and if p is of the form $3m+2$ then $[p/3, 2p/3] = \{m+1, \dots, 2m+1\}$.

In both cases, $[p/3, 2p/3]$ contains at least a third of the non-zero residues mod p. Therefore, for each $x \in X$, the probability that $ax \in [p/3, 2p/3]$ is at least $1/3$, so the expected number of $x \in X$ such that $ax \in [p/3, 2p/3]$ is at least $|X|/3$.

Applying the basic principle, there must exist a such that for at least a third of the elements x of X we have $ax \in [p/3, 2p/3]$. Let Y be the set of all such x. Then $|Y| \geq n/3$. But also Y is sum free, since if $x, y, z \in Y$ and $x + y = z$, then we would have $ax + ay \equiv az$ mod p, which is impossible because the set $[p/3, 2p/3]$ is sum free mod p. □

The difference between the way I presented the proof there and the way I described it in advance is that in the proof above I kept the interval $[p/3, 2p/3]$ fixed and multiplied X by a random a mod p, whereas in the description before it I kept X fixed and took multiples of the interval $[p/3, 2p/3]$. Thus, the two ideas are not different in any fundamental way.

Here is a summary of the above proof. Again, this is what you should remember rather than the proof itself.

(1) The "middle third" of the integers mod p form a sum-free set mod p.
(2) If we multiply everything in X by some number and reduce mod p, then we preserve all relationships of the form $x + y = z$.
(3) But if $p > \max X$ and we choose a randomly from non-zero integers mod p, then on average a third of the elements of X end up in the middle third.
(4) Therefore, we are done, by the basic averaging principle.

The third example is a beautiful proof of a result about crossing numbers. The *crossing number* of a graph G is the smallest possible number of pairs of edges that cross in any drawing of G in the plane. We would like a lower bound for this number in terms of the number of vertices and edges of G.

We begin with a well-known lemma that tells us when we are forced to have one crossing.

Lemma 2.3. *Let G be a graph with n vertices and more than $3n - 6$ edges. Then G is not planar. (In other words, for every drawing of G in the plane there must be at least one pair of edges that cross.)*

Proof. Euler's famous formula for drawings of planar graphs is that $V - E + F = 2$, where V is the number of vertices, E the number of edges and F the number of faces (including an external face that lies outside the entire graph).

Since each edge lies in two faces and each face has at least three edges on its boundary, we must have (by counting pairs (e, f) where e is an edge on the boundary of face f) that $2E \geq 3F$, so that $F \leq 2E/3$. It follows that $V - E/3 \geq 2$, so $E \leq 3V - 6$.

Therefore, if a graph with n vertices has more than $3n - 6$ edges, then it is not planar, as claimed. □

A famously non-planar graph is K_5, the complete graph with five vertices. That has $\binom{5}{2} = 10$ edges, and since $10 = 3 \times 5 - 5$, its non-planarity follows from the lemma above.

To avoid cumbersome sentences, I'm going to use the word "crossing" to mean "pair of edges that cross" rather than "point where two edges cross".

Corollary 2.4. *Let G be a graph with n vertices and m edges. Then the number of crossings is at least $m - 3n$.*

Proof. Informally, we just keep removing edges that are involved in crossings until we get down to $3n$ edges. More formally, the result is true when $m = 3n + 1$, by the previous lemma. But if it is true for $3n + k$, then for any graph with $3n + k + 1$ edges there is at least one edge that crosses another. Removing such an edge results in a graph with $3n + k$ edges and hence, by the inductive hypothesis, at least k crossings, which do not involve the removed edge. So the number of crossings is at least $k + 1$ and we are done. □

That didn't have much to do with averaging, but the averaging comes in at the next stage.

The proof of the corollary above feels a bit wasteful, because each time we remove an edge, we take account of only one crossing that involves that edge, when there could in principle be many. We shall now try to "boost" the result to obtain a much better bound when m is large.

Why might we expect that to be possible? The basic idea behind the next proof is that if G has lots of edges and we choose a random induced subgraph of G, then it will still have enough edges to force a crossing. The reason this helps is that (i) the random induced subgraphs "cover G evenly" in the sense discussed earlier, and (ii) if we choose p appropriately, then the previous corollary gives an efficient bound. So, roughly speaking, we pass to a subgraph where the argument we have above is not wasteful, and then use the

evenness of the covering to argue that G must have had lots of crossings for the random induced subgraph to have as many as it does.

Those remarks may seem a little cryptic: I recommend trying to understand them in conjunction with reading the formal proof.

For a graph G I'll use the notation $v(G)$ for the number of vertices and $e(G)$ for the number of edges.

Theorem 2.5. *Let G be a graph with n vertices and $m \geq 4n$ edges. Then G the crossing number of G is at least $m^3/64n^2$.*

Proof. Suppose we have a drawing of G with k crossings. We now choose a random induced subgraph H of G by selecting the vertices independently with probability $p = 4n/m$. (The reason for this choice of p will become clear later in the proof. For now we shall just call it p. Note that once the vertices of H are specified, so are the edges, since it is an induced subgraph.)

Then the expectation of $v(H)$ is pn, since each vertex survives with probability p, the expectation of $e(H)$ is p^2m, since each edge survives with probability p^2, and the expectation of the number of crossings of H (in this particular drawing) is p^4k, since each crossing survives with probability k.

But we also know that the expected number of crossings is at least $\mathbb{E}(e(H) - 3v(H))$, by the previous corollary. We have chosen p such that $e(H) - 3v(H) = p^2m - 3pn = 4pn - 3pn = pn$. Therefore, $p^4k \geq pn$, which implies that $k \geq p^{-3}n = m^3/64n^2$, as stated. □

It might be worth explaining why we chose p to be $4n/m$ rather than, say, $3n/m$. One reason is that if you do the calculations with $3n/m$, you get only six crossings in H (if you are slightly more careful with the corollary) instead of pn crossings, and therefore the answer you get out at the end is $6p^{-4}$, which is proportional to m^4/n^4. Except when m is proportional to n^2 (i.e., as big as it can be), this gives a strictly weaker bound.

But that is a slightly unsatisfactory answer, since it doesn't explain *why* the bound is weaker. The best reason I can come up with (but I think it could probably be improved) is that Corollary 2.4 does not start to become wasteful until m is superlinear in n, since if m is a multiple of n, then one normally expects an edge to be involved in only a bounded number of crossings. Since increasing p from $3m/n$ to $4m/n$ doesn't cost us much – it just changes an absolute constant – we might as well give ourselves a number of edges that grows linearly with $v(H)$ rather than remaining bounded.

I think my summary of this proof would simply be, "Pick a random induced subgraph with p just large enough for Corollary 2.4 to give a linear number of crossings, and use averaging."

The next result is another one where the basic idea is to find a family of sets that uniformly covers the structure we are interested in, and then use an averaging argument to allow ourselves to drop down to a member of that family. It is a famous result of Erdős, Ko and Rado. Given a set X and a positive integer k, write $X^{(k)}$ for the set of all subsets of X of size k. And given a family of sets $\mathcal{A}$, say that it is *intersecting* if $A \cap B \neq \emptyset$ whenever $A, B \in \mathcal{A}$.

Theorem 2.6. *Let X be a set of size n, let k be such that $2k \leq n$, and let $\mathcal{A} \subset X^{(k)}$ be an intersecting family. Then $|\mathcal{A}| \leq \binom{n-1}{k-1}$.*

Note that it is easy to see that this result is sharp, since we can take all sets of size k that contain one particular element. Also, the condition that $2k \leq n$ is obviously necessary, since otherwise we can take the entire set $X^{(k)}$.

Proof. The following very nice proof is due to Katona. Let $X = \{x_1, \dots, x_k\}$, and define an *interval* to be a set of the form $\{x_j, x_{j+1}, \dots, x_{j+k-1}\}$, where addition of the indices is mod n. There are n intervals, and if you want to form an intersecting family out of them, you can pick at most k. To prove this, let $\{x_j, \dots, x_{j+k-1}\}$ be one of the intervals in the family and observe that for any other interval J in the family there must exist $1 \leq h \leq k-1$ such that exactly one of x_{j+h-1} and x_{j+h} belongs to J, and for any h there can be at most one such interval J.

The rest of the proof, if you are used to the basic method, is now obvious. We understand families of intervals, so would like to cover $X^{(k)}$ evenly with them and see what we can deduce from the fact that $\mathcal{A}$ cannot have too large an intersection with any one of them. The obvious way of making this idea work is to choose a random ordering $\{x_1, \dots, x_n\}$ of the elements of X and take the intervals with respect to that ordering. The expected number of elements of $\mathcal{A}$ amongst these intervals is $n|\mathcal{A}|/\binom{n}{k}$, since each set in $\mathcal{A}$ has a $n/\binom{n}{k}$ chance of being an interval in this ordering (because there are n of them and $\binom{n}{k}$ sets of size k, and by symmetry each set of size k has an equal probability of being an interval in the ordering).

By the basic principle, there must therefore be an ordering such that at least $n|\mathcal{A}|/\binom{n}{k}$ sets in $\mathcal{A}$ are intervals. Since we also know that at most k intervals can all intersect each other, we must have that $n|\mathcal{A}|/\binom{n}{k} \leq k$, so $|\mathcal{A}| \leq \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}$. □

If you like thinking about semi-philosophical questions, then you could try to find a good explanation for why this argument gives the best possible bound.

Here is how I would remember the above proof.

(1) It is not too hard to show that amongst all the intervals of length k in a cyclic ordering of the elements of X, at most k can intersect.
(2) Hence, by averaging over all cyclic orderings, the density of $\mathcal{A}$ in $X^{(k)}$ is at most k/n.

Actually, that "it is not too hard to show" is slightly dishonest: I find that each time I come back to this theorem I have forgotten the argument that proves this "obvious" fact. But it is certainly a short argument and easy to understand.

For a final example, we shall give a proof of Sperner's theorem, which concerns the maximum possible size of an *antichain*: a collection of sets none of which is a proper subset of another.

Theorem 2.7. *Let X be a set of size n and let $\mathcal{A}$ be an antichain of subsets of X. Then $|\mathcal{A}| \leq \binom{n}{\lfloor n/2 \rfloor}$.*

Proof. Again we shall cover the structure we are interested in – the set of all subsets of X – with a family of sets. An interesting twist here is that the covering is *not* even. But as we shall see, this does not matter. In fact, it allows us to prove a stronger result than the one stated.

The preliminary observation that gets the proof going is that for any collection of sets of the form $\emptyset, \{x_1\}, \{x_1, x_2\}, \dots, \{x_1, x_2, \dots, x_n\}$, at most one of them can belong to $\mathcal{A}$. So our family will consist of sets of this kind, which are known as *maximal chains*.

Now we just do the obvious thing. Choose a random ordering $x_1, \dots, x_n$ of the elements of X and form the maximal chain $\emptyset, \{x_1\}, \{x_1, x_2\}, \dots, \{x_1, x_2, \dots, x_n\}$. If $A \in \mathcal{A}$, then the probability that it belongs to this chain is the probability that it equals the set $\{x_1, \dots, x_k\}$, where $k = |A|$. This probability is $\binom{n}{k}^{-1}$. Therefore, the expected number of sets in $\mathcal{A}$ that belong to the random maximal chain is $\sum_{A \in \mathcal{A}} \binom{n}{|A|}^{-1}$.

But it is also at most 1, since no maximal chain contains more than one set from $\mathcal{A}$. Since $\binom{n}{|A|} \leq \binom{n}{\lfloor n/2 \rfloor}$ regardless of the size of A, it follows that $|\mathcal{A}| \leq \binom{n}{\lfloor n/2 \rfloor}$, as stated. □

I didn't explicitly use the basic principle there, but I could have. I could have said that by the principle there exists a maximal chain that contains at least $\sum_{A \in \mathcal{A}} \binom{n}{|A|}^{-1}$ sets from $\mathcal{A}$. Then I would have concluded the proof from the fact that it also contains at most one set from $\mathcal{A}$.

This proof is another one with a very simple summary: pick a random maximal chain, calculate the expected number of sets in $\mathcal{A}$ that it contains, and draw the obvious conclusion.

I said that the proof gives a stronger result. What I meant by that can be stated as follows. Suppose we assign a measure of $\binom{n}{k}^{-1}$ to each set of size k. Then the total measure of $\mathcal{A}$ is, by the proof, at most 1. Since most sets have measure greater than $\binom{n}{\lfloor n/2\rfloor}^{-1}$, and some have measure much greater than this, the result with this non-uniform measure is significantly stronger.

3. Using mean and variance

So far we have seen that the inequality

$$\inf X \leq \mathbb{E}X \leq \sup X$$

is surprisingly useful. Now let's do the same for a well-known identity that is not quite as trivial, but is still very easy – so much so that it is again rather extraordinary how many applications it has. This time I'll give a proof.

Lemma 3.1. *Let X be a random variable. Then* $\mathrm{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

Proof. The variance $\mathrm{var}X$ is defined to be $\mathbb{E}(X - \mathbb{E}X)^2$. But

$$\mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - 2(\mathbb{E}X)\mathbb{E}X + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2,$$

which gives us the identity. □

While we're at it, here are a couple of useful inequalities. The first is *Markov's inequality*.

Lemma 3.2. *Let X be a random variable that takes non-negative real values. Then for any non-negative real number, $\mathbb{P}[X \geq a] \leq a^{-1}\mathbb{E}X$.*

Proof. Let us bound the expectation from below by partitioning the sample space into the set where $X \geq a$ and the set where $X < a$. Since $X \geq 0$ on the second set, we get

$$\mathbb{E}X \geq a\mathbb{P}[X \geq a],$$

from which the inequality follows. □

The second is *Chebyshev's inequality*.

Lemma 3.3. *Let X be a random variable and let $a \geq 0$. Then $\mathbb{P}[|X - \mathbb{E}X| \geq a] \leq a^{-2}\mathrm{var}X$.*

Proof. Let $Y = (X - \mathbb{E}X)^2$. Then $\mathbb{E}Y = \text{var}X$. Therefore, by Markov's inequality,

$$\mathbb{P}[|X - \mathbb{E}X| \geq a] = \mathbb{P}[Y \geq a^2] \leq a^{-2}\text{var}X,$$

which proves the result. □

Now let us use these simple facts to prove a not quite so simple result. Write $\mathbb{Z}_p$ for the set of integers mod p. Throughout the next result, addition is mod p.

Theorem 3.4. *Let A be a subset of $\mathbb{Z}_p$ with δp elements. Suppose that there are at most $\delta^4(1+\gamma)p^3$ quadruples $(a,b,c,d) \in A^4$ with $a+b=c+d$. Then there are at least $\delta^3(1-2\gamma^{1/3}\delta^{-1/3})p^2$ pairs (a,d) such that $(a,a+d,a+2d) \in A^3$.*

Proof. For each $x \in \mathbb{Z}_p$, let $f(x)$ be the number of ways of writing $x = a+b$ with $a, b \in A$. Equivalently, $f(x)$ is the number of $a \in A$ such that $x - a \in A$ as well.

The number of quadruples $(a,b,c,d) \in A^4$ such that $a+b=c+d=x$ is $f(x)^2$, since there are $f(x)$ ways of choosing a, b and $f(x)$ independent ways of choosing c, d. Therefore, we are given that $\sum_x f(x)^2 \leq \delta^4(1+\gamma)p^3$.

We also have that $\sum_x f(x) = |A|^2 = \delta^2 p^2$, since every pair $(a,b) \in A^2$ makes a contribution of 1 to the sum. (Another way of putting it is that $\sum_x f(x)$ is the number of ways of writing *some* number in $\mathbb{Z}_p$ as a sum of two elements a, b of A, taking their order into account, and that is clearly the number of ways of choosing a times the number of ways of choosing b.)

Let us think of $\mathbb{Z}_p$ as a sample space and f as a random variable. Then $\mathbb{E}f = p^{-1}\sum_x f(x) = \delta^2 p$, and $\mathbb{E}f^2 = p^{-1}\sum_x f(x)^2 \leq \delta^4(1+\gamma)p^2$. By Lemma 3.1, it follows that $\text{var}f \leq \gamma\delta^4 p^2$. Therefore, by Chebyshev's inequality,

$$\mathbb{P}[f(x) \leq \delta^2(1-\epsilon)p] \leq (\epsilon\delta^2 p)^{-2}\gamma\delta^4 p^2 = \gamma\epsilon^{-2}.$$

Now the number of pairs (a,d) such that $(a,a+d,a+2d) \in A^3$ is equal to the number of triples $(x,y,z) \in A^3$ such that $x+z=2y$, which equals $\sum_{y\in A} f(2y)$. If we pick y randomly, then the probability that it lies in A is δ and the probability that $f(2y) \leq \delta^2(1-\epsilon)p$ is at most $\gamma\epsilon^{-2}$. Therefore,

$$\mathbb{E}_y 1_A(y) f(2y) \geq \delta^2(1-\epsilon)p(\delta - \gamma\epsilon^{-2}) = \delta^3 p(1-\epsilon)(1-\gamma\epsilon^{-2}\delta^{-1}).$$

If we take $\epsilon = (\gamma\delta^{-1})^{1/3}$, this gives us $\delta^3 p(1-\gamma^{1/3}\delta^{-1/3})^2 \geq \delta^3 p(1-2\gamma^{1/3}\delta^{-1/3})$. The result follows on multiplying by p (to convert the expectation over y into a sum). □

It often happens in this area that one can do a slightly ugly argument using Markov's inequality or Chebyshev's inequality, but with a bit more effort one can find a more elegant argument that gives a better bound. Here we can do that by introducing the function $g(x) = f(x) - \delta^2 p$, which averages zero. We also know that $\sum_x g(x)^2 \leq \gamma\delta^4 p^3$, from which it follows that $\sum_{x\in A} g(2x)^2 \leq \gamma\delta^4 p^3$, and therefore $\mathbb{E}_{x\in A} g(2x)^2 \leq \gamma\delta^3 p^2$. Since the variance of g is non-negative, it follows that $(\mathbb{E}_{x\in A} g(2x))^2 \leq \gamma\delta^3 p^2$, so $\mathbb{E}_{x\in A} g(2x) \geq -\gamma^{1/2}\delta^{3/2} p$, and therefore $\sum_{x\in A} g(2x) \geq -\gamma^{1/2}\delta^{5/2} p^2$.

It follows that $\sum_{x\in A} f(2x) \geq (\delta^3 - \gamma^{1/2}\delta^{5/2})p^2 = \delta^3 p^2(1 - \gamma^{1/2}\delta^{-1/2})$.

Once again I recommend not remembering the proof (or proofs) above, but a summary such as this.

(1) The condition about the number of quadruples $a + b = c + d$ can be interpreted as an upper bound for the second moment of the function that tells you how many ways you can write x as $a + b$.
(2) The mean of this function is easy to determine, and it gives us that the variance is small.
(3) But if the variance is small enough, then the function is close to its mean almost everywhere, and in particular on most points of the form $2x$ for some $x \in X$.

We shall use second-moment methods quite a bit during this course. But to illustrate that they can come up in a wide variety of contexts, we now give an application to analytic number theory, namely a famous argument of Turán, which greatly simplified an earlier proof of the same result that had been more number-theoretic in character.

We shall need the following estimate for the sum of the reciprocals of the primes up to n. The proof will be left as an exercise (with hints) on an examples sheet. (It follows from the prime number theorem, but can be proved much more easily.)

Lemma 3.5. *There exists a positive integer C such that $\sum_{p\leq n} p^{-1} \leq \log\log n + C$ for every n, where the sum is over primes p.*

The theorem we shall now prove tells us that almost all numbers up to n have roughly $\log\log n$ prime factors.

Theorem 3.6. *Let x be a randomly chosen integer between 1 and n. Let $\nu(x)$ be the number of prime factors of x (without multiplicity). And let $\omega : \mathbb{N} \to \mathbb{N}$ be a function that tends to infinity. Then*

$$\mathbb{P}[|\nu(x) - \log\log n| > \omega(n)\sqrt{\log\log n}] = o(1).$$

Proof. To prove this, the rough idea is to show that the mean and variance of $\nu(x)$ are both approximately $\log\log n$. Then the probability that $|\nu(x) - \log\log n| \geq C\sqrt{\log\log n}$ will (if our approximation is good enough) be at most approximately C^{-2}, by Chebyshev's inequality.

To estimate the mean and variance, we write $\nu(x)$ as a sum of random variables that are easy to understand. For each prime p, let X_p be the random variable that takes the value 1 if $p|x$ and 0 otherwise. Then $\nu(x) = \sum_p X_p(x)$, where the sum is over all primes.

Let $m = n^{1/2}$. The reason I wrote "the rough idea" above is that it turns out to be convenient to estimate the mean and variance of $\nu_1(x) = \sum_{p\leq m} X_p(x)$ instead of those of ν. Since x can have at most one prime factor greater than m, we have $\nu_1(x) \leq \nu(x) \leq \nu_1(x)+1$ for every x. So the theorem for ν_1 implies the theorem for ν.

The mean of X_p is $n^{-1}\lfloor n/p\rfloor$, which lies between $1/p - 1/n$ and $1/p$. Therefore,

$$(\sum_{p\leq m}\frac{1}{p}) - 1 \leq \mathbb{E}\nu_1 \leq \sum_{p\leq m}\frac{1}{p}.$$

By Lemma 3.5, this is at most $\log\log n + C$.

We want to work out the variance of ν_1, so let us work out $\mathbb{E}\nu_1^2$ and use Lemma 3.1. We have

$$\mathbb{E}\nu_1^2 = \mathbb{E}(\sum_{p\leq m} X_p)^2 = \sum_{p\leq n}\mathbb{E}X_p + \sum_{p,q\leq m,\ p\neq q}\mathbb{E}X_pX_q.$$

We also have

$$(\mathbb{E}\nu_1)^2 = (\mathbb{E}\sum_{p\leq m} X_p)^2 = \sum_{p\leq m}(\mathbb{E}X_p)^2 + \sum_{p,q\leq m,\ p\neq q}(\mathbb{E}X_p)(\mathbb{E}X_q).$$

It follows that

$$\operatorname{var}\nu_1 \leq \log\log n + C + \sum_{p,q\leq m,\ p\neq q}(\mathbb{E}X_pX_q - \mathbb{E}X_p\mathbb{E}X_q).$$

Let us now estimate $\mathbb{E}X_pX_q - \mathbb{E}X_p\mathbb{E}X_q$, the covariance of X_p and X_q. It is at most

$$\frac{1}{pq} - (\frac{1}{p} - \frac{1}{n})(\frac{1}{q} - \frac{1}{n}) \leq \frac{1}{n}(\frac{1}{p} + \frac{1}{q}).$$

Adding these together for all $p, q \leq m$ gives us $2(m/n)\sum_{p\leq n} p^{-1}$, which is much smaller than 1. Therefore, we find that $\operatorname{var}\nu_1 \leq \log\log n + C + 1$ and the theorem follows. □

A famous theorem of Erdős and Kac states that the distribution of $\nu(x)$ is roughly normal with mean and variance $\log\log n$. Very roughly, the idea of the proof is to use an argument like the one above to show that every moment of ν (or more precisely a

distribution where, as with ν_1, one truncates the range of summation of the primes) is roughly equal to the corresponding moment for the normal distribution, which is known to imply that the distribution itself is roughly normal.

4. Using the Cauchy-Schwarz inequality

It is time to introduce some notation that has many virtues and has become standard in additive combinatorics. If X is a set of size n and f is a function on X taking values in $\mathbb{R}$ or $\mathbb{C}$, then we write $\mathbb{E}f$ or $\mathbb{E}_x f(x)$ for $n^{-1}\sum_{x\in X} f(x)$. For $1 \leq p < \infty$ we define the L_p norm of such a function f to be $(\mathbb{E}_x|f(x)|^p)^{1/p}$, and write it as $\|f\|_p$. We also define the L_∞ norm $\|f\|_\infty$ to be $\max_x |f(x)|$.

Typically, the L_p norm is useful when the function f is "flat", in the sense that its values mostly have the same order of magnitude. As we shall see later, we often like to set things up so that we are considering functions where the values have typical order of magnitude 1.

However, sometimes we also encounter functions for which a few values have order of magnitude 1, but most values are small. In this situation, it tends to be more convenient to use the uniform *counting* measure on X rather than the uniform probability measure. Accordingly, we define the ℓ_p norm of a function f defined on X to be $(\sum_{x\in X}|f(x)|^p)^{1/p}$. We denote this by $\|f\|_p$ as well, but if the meaning is unclear from the context, then we can always write $\|f\|_{L_p}$ and $\|f\|_{\ell_p}$ instead. Note that we also define the ℓ_∞ norm, but it is equal to the L_∞ norm. (That is because $n^{1/p} \to 1$ as $p \to \infty$, so the limit of the ℓ_p norms equals the limit of the L_p norms.)

As with norms, we also define two inner products. Given two functions $f, g : X \to \mathbb{C}$, these are given by the formulae $\mathbb{E}_x f(x)\overline{g(x)}$ and $\sum_x f(x)\overline{g(x)}$. Both are denoted by $\langle f, g\rangle$, and again the context makes clear which is intended. We have the identity $\langle f, f\rangle = \|f\|_2^2$, provided we either use expectations on both sides or sums on both sides.

Now let us have two lemmas. The first is the Cauchy-Schwarz inequality itself. Again, the result has two interpretations, both of which are valid, provided one is consistent about whether one is using sums or expectations.

Lemma 4.1. *Let X be a finite set and let f and g be a function defined on X that takes values in $\mathbb{R}$ or $\mathbb{C}$. Then $\langle f, g\rangle \leq \|f\|_2\|g\|_2$.*

The sum version will be familiar to you already. If we use expectations instead, then we have to divide the inner product by $|X|$ and the square of each norm by $|X|$, so the result still holds. The expectations version can also be thought of as an instance of the probability version of the inequality, which states that for two random variables X and Y

we have $\mathbb{E}XY \leq (\mathbb{E}|X|^2)^{1/2}(\mathbb{E}|Y|^2)^{1/2}$. In this case, the random variables are obtained by picking a random element of x and evaluating f and g. (Apologies for the overuse of the letter X here.)

The second lemma is another surprisingly useful tool.

Lemma 4.2. *Let X be a finite set and let f be a function from X to $\mathbb{R}$ or $\mathbb{C}$. Then*

$$|\mathbb{E}_x f(x)|^2 \leq \mathbb{E}_x |f(x)|^2.$$

Proof. If f is real valued, then $\mathbb{E}_x|f(x)|^2 - |\mathbb{E}_x f(x)|^2$ is $\mathrm{var}f$, by Lemma 3.1, and the variance is non-negative by definition.

To prove the result in the complex case, we simply check that an appropriate generalization of Lemma 3.1 holds, which indeed it does, since

$$\mathbb{E}_x|f(x) - \mathbb{E}f|^2 = \mathbb{E}_x\big(|f(x)|^2 - f(x)\overline{\mathbb{E}f} - \overline{f(x)}\mathbb{E}f + |\mathbb{E}f|^2\big) = \mathbb{E}_x|f(x)|^2 - |\mathbb{E}f|^2.$$

This proves the lemma. □

A second proof is to apply the Cauchy-Schwarz inequality to the function f and the constant function that takes the value 1 everywhere. That tells us that $|\mathbb{E}_x f(x)| \leq \|f\|_2$, since the constant function has L_2 norm 1, and squaring both sides gives the result. I myself prefer the proof given above, because it is simpler, and because by focusing attention on the variance it also makes another important fact very clear, which is that if $|\mathbb{E}_x f(x)|^2$ is almost as big as $\mathbb{E}_x|f(x)|^2$, then the variance of f is small, and therefore f is approximately constant. (The "approximately" here means that the difference between f and a constant function is small in the L_2 norm.) This very simple principle is yet another one with legions of applications. Indeed, we have already seen one: it was Theorem 3.4.

We have just encountered our first payoff for using expectations rather than sums. The sums version of Lemma 4.2 is that $|\sum_x f(x)|^2 \leq |X|\sum_x |f(x)|^2$. That is, we have to introduce a normalizing factor $|X|$ – the size of the set on which f takes its values. The great advantage of the expectation notation is that for many arguments it allows one to make all quantities one is interested in have order of magnitude 1, so we don't need any normalizing factors. This provides a very useful "checking of units" as our arguments proceed.

Having stated these results, let us begin with a simple, but very typical, application. Let us define a *4-cycle* in a graph G to be an ordered quadruple (x, y, z, w) of vertices such that all of xy, yz, zw, wx are edges. This is a non-standard definition because (i) we count (x, y, z, w) as the same 4-cycle as, for instance, (z, y, x, w), and (ii) we do not insist that

x, y, z, w are distinct. However, the proofs run much more smoothly with this non-standard definition, and from the results it is easy to deduce very similar results about 4-cycles as they are usually defined.

Theorem 4.3. *Let G be a graph with n vertices and $\alpha n^2/2$ edges. Then G contains at least $\alpha^4 n^4$ 4-cycles.*

Proof. The average degree of G is αn. Therefore, by Lemma 4.2 the average square degree is at least $\alpha^2 n^2$. So $\sum_{x \in V(G)} d(x)^2 \geq \alpha^2 n^3$.

This sum is the number of triples (x, y, z) such that xy and xz are both edges. Let us define the *codegree* $d(y, z)$ to be the number of vertices x that are joined to both y and z. Then the number of the triples (x, y, z) with $xy, xz \in E(G)$ is equal to $\sum_{y,z} d(y, z)$. Therefore, $\mathbb{E}_{y,z} d(y, z) \geq \alpha^2 n$. By Lemma 4.2 again, $\mathbb{E}_{y,z} d(y, z)^2 \geq \alpha^4 n^2$, so $\sum_{y,z} d(y, z)^2 \geq \alpha^4 n^4$.

But $\sum_{y,z} d(y, z)^2$ is the sum over all y, z of the number of pairs x, x' such that yx, zx, yx' and zx' are all edges. In other words it is the number of 4-cycles (y, x, z, x'). □

It was slightly clumsy to pass from sums to expectations and back to sums again in the above proof. But this clumsiness, rather than indicating that there is a problem with expectation notation, actually indicates that we did not go far enough. We can get rid of it by talking about the *density* of the graph, which is α, and proving that the 4-cycle density (which I will define it later, but it should be clear what I am talking about) is at least α^4.

Our next application will use the full Cauchy-Schwarz inequality. As well as illustrating an important basic technique, it will yield a result that will be important to us later on. We begin by defining a norm on functions of two variables. We shall apply it to real-valued functions, but we shall give the result for complex-valued functions, since they sometimes come up in applications as well.

Let X and Y be finite sets and let $f : X \times Y \to \mathbb{C}$. The *4-cycle norm* $\|f\|_\square$ is defined by the formula

$$\|f\|^4 = \mathbb{E}_{x_0, x_1, y_0, y_1 \in X} f(x_0, y_0) \overline{f(x_0, y_1) f(x_1, y_0)} f(x_1, y_1).$$

The proof that this formula defines a norm is another application of the Cauchy-Schwarz inequality and will be presented as an exercise on an examples sheet. Here we shall prove an important inequality that tells us that functions with small 4-cycle norm have small correlation with functions of rank 1. I shall include rather more chat in the proof than is

customary, since I want to convey not just the proof but the general method that can be used to prove many similar results.

Theorem 4.4. *Let X and Y be finite sets, let $u : X \to \mathbb{C}$, let $v : Y \to \mathbb{C}$ and let $f : X \to \mathbb{C}$. Then*

$$|\mathbb{E}_{x,y} f(x,y)u(x)v(y)| \leq \|f\|_\square \|u\|_2 \|v\|_2.$$

Proof. With this sort of proof it is usually nicer to square both sides of the Cauchy-Schwarz inequality, so let us look at the quantity $|\mathbb{E}_{x,y} f(x,y)u(x)v(y)|^2$.

The basic technique, which is applied over and over again in additive combinatorics, is to "pull out" a variable or variables, apply the Cauchy-Schwarz inequality, and expand out any squared quantities. Here is the technique in action. We shall first pull out the variable x, which means splitting the expectation into an expectation over x and an expectation over y, with everything that depends just on x pulled outside the expectation over y. In symbols,

$$\mathbb{E}_{x,y} f(x,y)u(x)v(y) = \mathbb{E}_x u(x)\mathbb{E}_y f(x,y)v(y).$$

Thus, the quantity we want to bound is $|\mathbb{E}_x u(x)\mathbb{E}_y f(x,y)v(y)|^2$.

Note that $\mathbb{E}_y f(x,y)v(y)$ is a function of x. Therefore, by the (squared) Cauchy-Schwarz inequality,

$$|\mathbb{E}_x u(x)\mathbb{E}_y f(x,y)v(y)|^2 \leq \|u\|_2^2 \,\mathbb{E}_x |\mathbb{E}_y f(x,y)v(y)|^2.$$

Now comes the third stage: expanding out the squared quantity. We have that

$$\mathbb{E}_x |\mathbb{E}_y f(x,y)v(y)|^2 = \mathbb{E}_x \mathbb{E}_{y_0,y_1} f(x,y_0)\overline{f(x,y_1)}v(y_0)\overline{v(y_1)}.$$

If you do not find that equality obvious, then you should introduce an intermediate step, where the modulus squared is replaced by the product of the expectation over y with its complex conjugate. But with a small amount of practice, these expansions become second nature.

Now we shall repeat the whole process in order to find an upper bound for

$$|\mathbb{E}_x \mathbb{E}_{y_0,y_1} f(x,y_0)\overline{f(x,y_1)}v(y_0)\overline{v(y_1)}|.$$

Pulling out y_0 and y_1 gives

$$|\mathbb{E}_x \mathbb{E}_{y_0,y_1} f(x,y_0)\overline{f(x,y_1)}v(y_0)\overline{v(y_1)}| = |\mathbb{E}_{y_0,y_1} v(y_0)\overline{v(y_1)}\mathbb{E}_x f(x,y_0)\overline{f(x,y_1)}|.$$

Then Cauchy-Schwarz gives

$$|\mathbb{E}_{y_0,y_1} v(y_0)\overline{v(y_1)}\mathbb{E}_x f(x,y_0)\overline{f(x,y_1)}| \leq (\mathbb{E}_{y_0,y_1}|v(y_0)\overline{v(y_1)}|^2)(\mathbb{E}_{y_0,y_1}|\mathbb{E}_x f(x,y_0)\overline{f(x,y_1)}|^2).$$

The first bracket on the right-hand side equals $(\mathbb{E}_y|v(y)|^2)^2 = \|v\|_2^4$. Expanding the second gives

$$\mathbb{E}_{y_0,y_1}\mathbb{E}_{x_0,x_1}f(x_0,y_0)\overline{f(x_0,y_1)f(x_1,y_0)}f(x_1,y_1),$$

which equals $\|f\|_\square^4$.

Putting all this together gives us that

$$|\mathbb{E}_{x,y}f(x,y)u(x)v(y)|^4 \leq \|f\|_\square^4\|u\|_2^4\|v\|_2^4,$$

which proves the result. □

5. A brief introduction to quasirandom graphs

A central concept in extremal combinatorics, and also in additive combinatorics, is that of *quasirandomness*. We have already seen that randomly chosen objects have some nice properties. In the late 1980s it was discovered that many of the properties exhibited by random graphs are, in a loose sense, equivalent. That is, if you have a graph with one "random-like" property, then it automatically has several other such properties. Graphs with one, and hence all, of these properties came to be called quasirandom. A little later, it was realized that the quasirandomness phenomenon (of several different randomness properties being equivalent) applied to many other important combinatorial objects, and also gave us an extremely useful conceptual tool for attacking combinatorial problems.

Perhaps the main reason quasirandomness is useful is the same as the reason that any non-trivial equivalence is useful: we often want to apply one property, but a different, equivalent, property is much easier to verify. I will discuss this in more detail when I have actually presented some equivalences.

I would also like to draw attention to a device that I shall use in this section, which is another very important part of the combinatorial armoury: analysing sets by looking at their characteristic functions. Typically, one begins by wanting to prove a result about a combinatorial object such as a bipartite graph, which can be thought of as a function from a set $X \times Y$ to the set $\{0,1\}$, and ends up proving a more general result about functions from $X \times Y$ to the closed interval $[0,1]$, or even to the unit disc in $\mathbb{C}$.

The characteristic function of a set A is often denoted by χ_A or 1_A. My preference is simply to denote it by A. That is, $A(x) = 1$ if $x \in A$ and 0 otherwise, and similarly for functions of more than one variable.

The next result is about functions, but it will soon be applied to prove facts about graphs. We shall not need the complex version in this course, but we give it here since it is only slightly harder than the real version.

Lemma 5.1. *Let X and Y be finite sets, let $f : X \times Y \to \mathbb{C}$, and let $\mathbb{E}_{x,y} f(x,y) = \theta$. For each x, y, define $g(x)$ to be $\mathbb{E}_y f(x,y)$ and $f'(x,y)$ to be $f(x,y) - g(x)$, and for each x define $g'(x)$ to be $g(x) - \theta$. Then*

$$\|f\|_\square^4 \geq \|f'\|_\square^4 + \|g'\|_2^4 + |\theta|^4.$$

Proof. Note first that

$$\begin{aligned} \|f\|_\square^4 &= \mathbb{E}_{x,x'} |\mathbb{E}_y f(x,y)\overline{f(x',y)}|^2 \\ &= \mathbb{E}_{x,x'} |\mathbb{E}_y (f'(x,y) + g(x))(\overline{f'(x',y) + g(x')})|^2. \end{aligned}$$

Since for each x we have $\mathbb{E}_y f'(x,y) = 0$, this is equal to

$$\mathbb{E}_{x,x'} |g(x)\overline{g(x')} + \mathbb{E}_y f'(x,y)\overline{f'(x',y)}|^2.$$

When we expand the square, the off-diagonal term is

$$2\Re\, \mathbb{E}_{x,x'} g(x)\overline{g(x')}\mathbb{E}_y \overline{f'(x,y)} f'(x',y) = 2\Re\, \mathbb{E}_y |\mathbb{E}_x g(x)\overline{f'(x,y)}|^2 \geq 0.$$

It follows that

$$\|f\|_\square^4 \geq \mathbb{E}_{x,x'} |g(x)|^2 |g(x')|^2 + \mathbb{E}_{x,x'} |\mathbb{E}_y f'(x,y)\overline{f'(x',y)}|^2 = \|g\|_2^4 + \|f'\|_\square^4.$$

But $\|g\|_2^2 = |\theta|^2 + \|g'\|_2^2$, so $\|g\|_2^4 \geq |\theta|^4 + \|g'\|_2^4$, so the result is proved. □

Given a bipartite graph G with finite vertex sets X and Y, define the *density* of G to be $e(G)/|X||Y|$ – that is, the number of edges in G divided by the number of edges in the complete bipartite graph with vertex sets X and Y. Given that we want to focus on quantities with order of magnitude 1, we shall prefer talking about densities of graphs to talking about the number of edges they have.

For a similar reason, let us define the *4-cycle density* of G to be the number of 4-cycles in G divided by $|X|^2|Y|^2$. This is equal to

$$\mathbb{E}_{x,x' \in X} \mathbb{E}_{y,y' \in Y} G(x,y)G(x,y')G(x',y)G(x',y') = \|G\|_\square^4,$$

where I have used the letter G to denote the characteristic function of the graph G.

Corollary 5.2. *Let X and Y be finite sets and let G be a bipartite graph with vertex sets X and Y and density δ. Suppose that the 4-cycle density of G is at most $\delta^4(1+c^4)$. Then $\|G-\delta\|_\square \leq 2c\delta$.*

Proof. Applying Lemma 5.1 with $f=G$ and $\theta=\delta$, we deduce that $\|f'\|_\square^4+\|g'\|_2^4\leq\delta^4c^4$, where $f'(x,y)=G(x,y)-\mathbb{E}_z f(x,z)$ and $g'(x)=\mathbb{E}_z f(x,z)-\delta$. This follows because the 4-cycle density assumption is equivalent to the statement that $\|G\|_\square^4\leq\delta^4(1+c^4)$.

If we regard g' as a function of both x and y that happens not to depend on y, then we find that

$$\|g'\|_\square^4=\mathbb{E}_{x,x'}g'(x)\overline{g'(x)g'(x')}g'(x')=\|g'\|_2^4.$$

It follows that

$$\|G-\delta\|_\square=\|f'+g'\|_\square\leq\|f'\|_\square+\|g'\|_\square\leq 2c\delta,$$

as claimed. □

We now come to what is perhaps the most important (but certainly not the only important) fact about quasirandom graphs, which follows easily from what we have already proved.

Given a bipartite graph G as above, and subsets $A\subset X$ and $B\subset Y$, define the *edge weight* $\eta(A,B)$ to be the number of edges from A to B divided by $|X||Y|$. Although we shall not need it immediately, it is worth also mentioning the *edge density* $d(A,B)$, which is the number of edges from A to B divided by $|A||B|$.

Theorem 5.3. *Let X and Y be finite sets and let G be a bipartite graph with vertex sets X and Y. Let the density of G be δ and suppose that the 4-cycle density of G is at most $\delta^4(1+c^4)$. Let $A\subset X$ and $B\subset Y$ be subsets of density α and β, respectively. Then*

$$|\eta(A,B)-\alpha\beta\delta|\leq 2c\delta(\alpha\beta)^{1/2}.$$

Proof. The edge weight $\eta(A,B)$ is equal to $\mathbb{E}_{x,y}G(x,y)A(x)B(y)$. Therefore,

$$|\eta(A,B)-\alpha\beta\delta|=|\mathbb{E}_{x,y}(G(x,y)-\delta)A(x)B(y)|.$$

By Lemma 4.4, the right-hand side is at most $\|G-\delta\|_\square\|A\|_2\|B\|_2$. Corollary 5.2 gives us that $\|G-\delta\|_\square\leq 2c\delta$, while $\|A\|_2$ and $\|B\|_2$ are equal to $\alpha^{1/2}$ and $\beta^{1/2}$, respectively. This proves the result. □

Now let us try to understand what the theorem above is saying. If you are not yet used to densities rather than cardinalities, then you may prefer to reinterpret it in terms of the latter. Then it is saying that the number of edges from A to B is, when c is small, close

to $\alpha\beta\delta|X||Y| = \delta|A||B|$. Now if we had chosen our graph *randomly* with edge-probability δ, then we would have expected the number of edges from A to B to be roughly $\delta|A||B|$. So Theorem 5.3 is telling us that if a graph does not contain many 4-cycles (recall that Theorem 4.3 states that the 4-cycle density must be at least δ^4 and we are assuming that it is not much more than that), then for any two sets A and B the number of edges between A and B is roughly what you would expect. Let us say in this case that the graph has *low discrepancy*.

An important remark here is that if we restate the result in terms of the density $d(A,B)$ then it says that $|d(A,B) - \delta| \leq 2c\delta(\alpha\beta)^{-1/2}$. Therefore, for fixed c the approximation becomes steadily worse as the densities of A and B become smaller. This is a problem for some applications, but often the sets A and B that we are interested in have a density that is bounded below – that is, independent of the sizes of X and Y.

I like to think of Theorem 5.3 as a kind of local-to-global principle. We start with a "local" condition – counting the number of 4-cycles – and deduce from it a "global" principle – that the edges are spread about in such a way that there are approximately the right number joining any two large sets A and B.

To explain that slightly differently, consider how one might try to verify the "global" statement. If you try to verify it directly, you find yourself looking at exponentially many pairs of sets (A,B), which is hopelessly inefficient. But it turns out that you can verify it by counting 4-cycles, of which there are at most $|X|^2|Y|^2$, which is far smaller than exponential. While this algorithmic consideration will not be our main concern, it is another reflection of the fact that the implication is not just some trivial manipulation but is actually telling us something interesting. And as it happens, there are contexts where the algorithmic considerations are important too, though counting 4-cycles turns out not to be the best algorithm for checking quasirandomness.

Before we go any further, let us prove an approximate converse to Theorem 5.3. For simplicity, we shall prove it for regular graphs only and leave the general case as an exercise.

Theorem 5.4. *Let X, Y and G be as in Theorem 5.3, let each vertex in X have degree $\delta|Y|$, let each vertex in Y have degree $\delta|X|$, and suppose that the 4-cycle density of G is at least $\delta^4(1+c)$. Then there exist sets $A \subset X$ and $B \subset Y$ of density α and β such that $\eta(A,B) \geq \alpha\beta\delta(1+c)$.*

Proof. Our assumption about the 4-cycle density is that

$$\mathbb{E}_{x,x' \in X, y,y' \in Y} G(x,y)G(x,y')G(x',y)G(x',y') \geq \delta^4(1+c).$$

From this it follows that there exist vertices $x' \in X$ and $y' \in Y$ such that

$$\mathbb{E}_{x\in X,y\in Y}G(x,y)G(x,y')G(x',y) \geq \delta^3(1+c),$$

since if a random pair x', y' is chosen, the probability that $G(x', y')$ is non-zero is δ. Let A be the set of x such that $G(x, y') \neq 0$ and let B be the set of y such that $G(x', y) \neq 0$. In other words, A and B are the neighbourhoods of y' and x', respectively. Then A and B have density δ, by our regularity assumption, while

$$\eta(A,B) = \mathbb{E}_{x\in X,y\in Y}G(x,y)A(x)B(y) \geq \delta^3(1+c).$$

This proves the result. □

To adapt this argument to a proof for non-regular graphs, one first shows that unless almost all degrees are approximately the same, there exist sets A and B for which $\eta(A, B)$ is significantly larger than $\alpha\beta\delta$. Then one shows that if almost all degrees *are* approximately equal, then the argument above works, but with a small error introduced that slightly weakens the final result.

In case it is not already clear, I call a bipartite graph of density δ quasirandom if its 4-cycle density is at most $\delta^4(1 + c)$ for some small c, which is equivalent to saying that $\eta(A, B)$ is close to $\delta\alpha\beta$ for all sets $A \subset X$ and $B \subset Y$ of density α and β, respectively.

If we want to talk about *graphs*, as opposed to bipartite graphs, we can simply take a graph G with vertex set X, turn it into a bipartite graph by taking two copies of x and joining x in one copy to y in the other if and only if xy is an edge of G, and use the definitions and results from the bipartite case. In particular, the 4-cycle density of G is simply $|V(G)|^{-4}$ times the number of quadruples (x, y, z, w) such that xy, yz, zw and wx are all edges of G, and G is quasirandom of density δ if the average degree is $\delta|V(G)|$ and the 4-cycle density is $\delta^4(1 + c)$ for some small c.

However, a slight difference between graphs and bipartite graphs is that we can talk about the density of edges within a subset. That is, we define $d(A)$ to be $|A|^{-2}$ times the number of pairs $(x, y) \in A^2$ such that xy is an edge of G. We can also talk about the weight $\eta(A)$ of the edges in A, which I prefer to write analytically as

$$\mathbb{E}_{x,y}A(x)A(y)G(x,y).$$

It is a straightforward exercise to prove that if $\eta(A)$ is always close to $\delta\alpha^2$ (where α is the density of A), then $\eta(A, B)$ is always close to $\delta\alpha\beta$ (where β is the density of B). The converse is of course trivial.

In the exercises, you will also see that there is an equivalent formulation of quasirandomness in terms of eigenvalues of the adjacency matrix of G. The natural matrix to associate with a bipartite graph is not the adjacency matrix but the matrix $G(x,y)$ where x ranges over X and y ranges over Y. This matrix need not be symmetric, or even square, so we don't get an orthonormal basis of eigenvectors. Instead we need to use the singular value decomposition. I do not plan to go into this in the course, though may perhaps set it as an exercise.

One of the most striking facts about quasirandom graphs is that they contain the "right" number of any small subgraph. I shall prove this for triangles in tripartite graphs and leave the more general result as an exercise.

Given a graph G and two (not necessarily disjoint) sets A and B of its vertices, write $G(A,B)$ for the bipartite graph with vertex sets A and B where $a \in A$ is joined to $b \in B$ if and only if ab is an edge of G.

Theorem 5.5. *Let G be a tripartite graph with vertex sets X, Y, Z. Let the densities of $G(X,Y)$, $G(Y,Z)$ and $G(X,Z)$ be α, β and γ, respectively. Suppose that the 4-cycle densities of $G(X,Y)$, $G(Y,Z)$ and $G(X,Z)$ are at most $\alpha^4(1+c^4)$, $\beta^4(1+c^4)$ and $\gamma^4(1+c^4)$. Then the triangle density differs from $\alpha\beta\gamma$ by at most $4c(\alpha\beta\gamma)^{1/2}$.*

Proof. The quantity we wish to estimate is

$$\mathbb{E}_{x,y,z} G(x,y)G(y,z)G(x,z).$$

Let us write $G(x,y) = f(x,y) + \alpha$, $G(y,z) = g(y,z) + \beta$ and $G(x,z) = h(x,z) + \gamma$. Then

$$\mathbb{E}_{x.y.z} G(x,y)G(y,z)G(x,z) - \alpha\beta\gamma$$

can be expanded as

$$\mathbb{E}_{x,y,z}(f(x,y)G(y,z)G(x,z) + \alpha g(y,z)G(x,z) + \alpha\beta h(x,z)).$$

By Corollary 5.2, $\|f\|_\square \leq 2c\alpha$. Therefore, for each fixed z, we have

$$|\mathbb{E}_{x,y} f(x,y)G(y,z)G(x,z)| \leq 2c\alpha(\mathbb{E}_x G(x,z))^{1/2}(\mathbb{E}_y G(y,z))^{1/2},$$

by Theorem 4.4. Taking the expectation over z, we get that

$$\begin{aligned}|\mathbb{E}_{x,y,z}f(x,y)G(y,z)G(x,z)| &\leq 2c\alpha\mathbb{E}_z(\mathbb{E}_xG(x,z))^{1/2}(\mathbb{E}_yG(y,z))^{1/2}\\ &\leq 2c\alpha(\mathbb{E}_z\mathbb{E}_xG(x,z))^{1/2}(\mathbb{E}_z\mathbb{E}_yG(y,z))^{1/2}\\ &= 2c\alpha(\beta\gamma)^{1/2},\end{aligned}$$

where the second inequality was Cauchy-Schwarz.

Essentially the same argument shows that

$$|\alpha\mathbb{E}_{x.y.z}g(y,z)G(x,z)| \leq 2c\alpha\beta\gamma^{1/2},$$

while the third term is zero. This gives us the bound stated (and in fact, a better bound, but one that is uglier to state and not of obvious importance). □

I cannot stress strongly enough that it is the basic method of proof that matters above, and not the precise details of its implementation. So let me highlight what I regard as the basic method, and then indicate how it could have been implemented differently.

(1) A step that is useful all over additive combinatorics is to decompose a function F that averages δ into two parts $f+\delta$, so that we can deal with the constant function δ, which we often regard as the main term, and the function f, which averages zero and which in many contexts is a kind of "error". (The word "error" is reasonable only if the function is small in some sense. Here the sense in question is that of having a small 4-cycle norm.)
(2) Having done that, one breaks up the original expression into terms, and attempts to show that all terms that involve the error functions are small.

In the proof above, I chose a slightly non-obvious (but often quite useful) way of implementing step 2. A more obvious way of doing it would simply have been to split the expression

$$\mathbb{E}_{x,y,z}(\alpha+f(x,y))(\beta+g(y,z))(\gamma+h(x,z))$$

into eight terms. The main term would again be $\alpha\beta\gamma$ and the remaining seven terms (three of which are zero) can be estimated using Theorem 4.4 as in the proof above.

A second remark I would like to make is that if we use the low-discrepancy property of quasirandom graphs instead, we get an easy combinatorial argument that again shows that the number of triangles is roughly as expected. Here is a quick sketch.

A preliminary observation, which we have already mentioned, is that if a graph has low discrepancy, then almost all degrees are roughly the same. We now pick a random vertex $x \in X$. Then with high probability its neighbourhoods in Y and Z have densities roughly α and γ, respectively. When that happens, then by the low-discrepancy property in $G(Y,Z)$, the edge weight between these two neighbourhoods is approximately $\alpha\beta\gamma$. This shows that for almost all $x \in X$, the probability that a random triangle containing x belongs to the graph is roughly $\alpha\beta\gamma$, and from that the result follows.

Note that we did not use the low discrepancy of the graphs $G(X,Y)$ and $G(X,Z)$ there, but simply the fact that they are approximately regular. If you look at the proof of Theorem 5.5 above, you will see that a similar remark applies: to bound the first term we needed $\|f\|_\square$ to be small, but to bound the second, it would have been sufficient to know that $\mathbb{E}_y g(y,z)$ was small for almost all z, which is the same as saying that for almost all z the density of the neighbourhood of z in Y is roughly $\beta|Y|$. (One can therefore go further – in both arguments – and say that what matters is that one of the three bipartite graphs should be quasirandom and one should be approximately regular on one side. For instance, in the discrepancy argument sketched above, if $G(Y,Z)$ has density α and low discrepancy, and almost all vertices in x have approximately $\gamma|Z|$ neighbours in Z, then writing $\delta_Y(x)$ for the density of the neighbourhood of x in Y, we have that for almost all x the density of triangles containing x is approximately $\beta\gamma\delta_Y(x)$, so the triangle density is approximately $\beta\gamma\mathbb{E}_x\delta_Y(x) = \alpha\beta\gamma$.)

A third remark is that if we are dealing with a graph rather than a tripartite graph, we can prove a similar result by making it tripartite in a similar way to the way we made it bipartite earlier. That is, we take three copies of the vertex set and join vertices according to whether the corresponding vertices are joined in the original graph.

As mentioned already, the above proofs work with obvious modifications to show that quasirandom graphs contain the right number of copies of any fixed graph. I described the result as striking: what is striking about it is that the converse implication is completely trivial: if you have the right number of any fixed graph, then in particular you have the right number of 4-cycles.

Finally, I should mention that there are close connections between the quasirandomness of a graph and the sizes of the eigenvalues of its adjacency matrix. This is an important topic, which will be covered (or at least introduced) in the examples sheets.

6. Discrete Fourier analysis and quasirandom sets

Let $n \geq 2$ be a positive integer and let $\mathbb{Z}_n$ be the group of integers mod n. Let $\omega = \exp(2\pi i/n)$, so that $\omega^n = 1$.

Given a function $f : \mathbb{Z}_n \to \mathbb{C}$ we define its *(discrete) Fourier transform* by the formula

$$\hat{f}(r) = \mathbb{E}_x f(x)\omega^{-rx}.$$

Note that this is the inner product of f with the function $x \mapsto \omega^{rx}$. One of the important facts about the functions ω^{rx} is that they form an orthonormal basis with respect to this inner product. Indeed, the inner product of the functions ω^{rx} and ω^{sx} is $\mathbb{E}_x\omega^{(r-s)x}$. If $r = s$, then this is 1, while if $r \neq s$, then by the formula for a geometric progression it is $(1 - \omega^{(r-s)n})/n(1 - \omega^{r-s}) = 0$.

There are three basic facts about the Fourier transform that are used over and over again.

Lemma 6.1. (The inversion formula.) *Let $f : \mathbb{Z}_n \to \mathbb{C}$. Then for every $x \in \mathbb{Z}_n$ we have*

$$f(x) = \sum_r \hat{f}(r)\omega^{rx}.$$

Proof. Let me justify this in two different ways. The first is simply to expand out $\hat{f}(r)$. That gives us that

$$\begin{aligned} \sum_r \hat{f}(r)\omega^{rx} &= \sum_r \mathbb{E}_y f(y)\omega^{-ry}\omega^{rx} \\ &= \mathbb{E}_y f(y) \sum_r \omega^{r(x-y)}. \end{aligned}$$

If $y = x$, then $\sum_r \omega^{r(x-y)} = n$, and otherwise it is 0. Also, the probability that $y = x$ if y is chosen randomly is $1/n$. Therefore, the last expression is equal to $f(x)$.

The second justification is simply to say that if $u_1, \dots, u_n$ is an orthonormal basis of a complex inner product space, and v is any vector in that space, then $v = \sum_r \langle v, u_r\rangle u_r$. If we take the orthonormal basis consisting of the functions ω^{rx} and $v = f$, then we get precisely the inversion formula. □

With a little thought, one can see that these two justifications are essentially the same argument. It's just that in the first argument we basically reprove the fact that the functions ω^{rx} form an orthonormal basis rather than using it.

Note that in the above result, the right-hand side involves a sum over Fourier coefficients. That is typical in discrete Fourier analysis: we take expectations when we are in "physical space" and sums when we are in "frequency space".

The next result is sometimes attributed to Parseval and sometimes to Plancherel. It is sometimes called a formula, sometimes an identity and sometimes a theorem. So it has at least six possible names – Wikipedia gives some indication about which name is appropriate in which context, if you care about these things.

Lemma 6.2. *Let f and g be functions from $\mathbb{Z}_n$ to $\mathbb{C}$. Then $\langle \hat{f}, \hat{g} \rangle = \langle f, g \rangle$. In particular, if $f : \mathbb{Z}_n \to \mathbb{C}$, then $\|\hat{f}\|_{\ell_2} = \|f\|_{L_2}$.*

Proof. This follows immediately from the fact that the functions ω^{rx} form an orthonormal basis and the Fourier coefficients $\hat{f}(r)$ are the coefficients of f with respect to this basis. However, it is instructive to see a calculational proof as well as this more theoretical argument. We have

$$\begin{aligned} \langle \hat{f}, \hat{g} \rangle &= \sum_r \hat{f}(r)\overline{\hat{g}(r)} \\ &= \sum_r \mathbb{E}_{x,y} f(x)\overline{f(y)}\omega^{-r(x-y)} \\ &= \mathbb{E}_{x,y} f(x)\overline{f(y)} \sum_r \omega^{-r(x-y)}. \end{aligned}$$

If $x = y$ then the sum over r is n, while if $x \neq y$ it is zero. Also, for each x the probability that $x = y$ is n^{-1}. Therefore, the right-hand side works out to be $\mathbb{E}_x f(x)\overline{g(x)} = \langle f, g \rangle$. □

The next result is the *convolution identity*, the result that sets the Fourier transform apart from just any old unitary map. Given two functions $f, g : \mathbb{Z}_n \to \mathbb{C}$, define their *convolution* $f * g$ by the formula

$$f * g(x) = \mathbb{E}_{y+z=x} f(y)g(z).$$

Occasionally one also wants to convolve two functions when the appropriate normalization uses the counting measure on $\mathbb{Z}_n$ rather than the uniform probability measure. Then the convolution is defined using sums. For instance, we might want to say

$$\hat{f} * \hat{g}(r) = \sum_{s+t=r} \hat{f}(s)\hat{g}(t).$$

In these notes, the convolution will always be defined using expectations unless we clearly state otherwise.

Lemma 6.3. *Let f and g be functions from $\mathbb{Z}_n$ to $\mathbb{C}$. Then for every $r \in \mathbb{Z}_n$,*

$$\widehat{f * g}(r) = \hat{f}(r)\hat{g}(r).$$

Proof. This we shall prove in a purely calculational way.

$$\begin{aligned}
\widehat{f * g}(r) &= \mathbb{E}_x f * g(x)\omega^{-rx} \\
&= \mathbb{E}_x \mathbb{E}_{y+z=x} f(y)g(z)\omega^{-rx} \\
&= \mathbb{E}_x \mathbb{E}_{y+z=x} f(y)g(z)\omega^{-r(y+z)} \\
&= \mathbb{E}_x \mathbb{E}_{y+z=x} (f(y)\omega^{-ry})(g(z)\omega^{-rz}) \\
&= \mathbb{E}_y \mathbb{E}_z (f(y)\omega^{-ry})(g(z)\omega^{-rz}) \\
&= \hat{f}(r)\hat{g}(r),
\end{aligned}$$

as required. □

There are a number of identities that can be proved either by direct calculation or, in most cases, simply by using the three basic facts above. I will give one example (just the theoretical argument).

Proposition 6.4. *Let $f : \mathbb{Z}_n \to \mathbb{C}$. Then*

$$\|\hat{f}\|_4^4 = \mathbb{E}_{x,a,b} f(x)\overline{f(x+a)f(x+b)}f(x+a+b).$$

Proof. There is a one-to-one correspondence between quadruples of the form $(x, x+a, x+b, x+a+b)$ and quadruples (x, y, z, w) such that $x + w = y + z$. Therefore, the right-hand side can be rewritten

$$\mathbb{E}_{x+w=y+z} f(x)f(w)\overline{f(y)f(z)} = \mathbb{E}_u \mathbb{E}_{x+w=y+z=u} f(x)f(w)\overline{f(y)f(z)}.$$

But this is $\langle f * f, f * f\rangle$, so by Parseval's identity and the convolution identity it equals $\langle \hat{f}^2, \hat{f}^2\rangle$, which equals $\sum_r |\hat{f}(r)|^4$, which is $\|\hat{f}\|_4^4$. □

If f is the characteristic function of a set A, then there are a couple more Fourier-analytic facts that we can add to our repertoire.

Lemma 6.5. *Let $A \subset \mathbb{Z}_n$ be a set of density α. Then $\hat{A}(0) = \alpha$ and $\sum_r |\hat{A}(r)|^2 = \alpha$. Also, $\hat{A}(-r) = \overline{\hat{A}(r)}$ for every r.*

Proof. The first equality is direct from the definition and the second follows immediately from Parseval's identity. The third applies to all real-valued functions f: it follows from the definition and the fact that $f(x)\omega^{-u} = \overline{f(x)\omega^u}$ for every x and u. □

It follows that if A has density α, then $\|\hat{A}\|_4^4 = \sum_r |\hat{A}(r)|^4 \geq \alpha^4$. This fact can be proved in at least three other ways. One is to use Proposition 6.4 and the Cauchy-Schwarz inequality, another is to define an associated graph and use the results of exercise 9 on sheet 2 and exercise 7 on sheet 3, and a third is to use the same associated graph but work directly with 4-cycles. (It is not important to have all these proofs of one simple fact, but if you understand how all the arguments are connected, then it greatly improves your understanding of the whole area.)

We saw in Theorem 3.4 that if n is odd and A has not many more than $\alpha^4 n^3$ quadruples of the form $x + y = z + w$, then it contains approximately $\alpha^3 n^2$ triples of the form $(a, a + d, a + 2d)$. [As the notes currently stand, that is not precisely what Theorem 3.4 states, but it is easy to see that the proof gives this stronger statement.] We shall now prove a closely related statement using Fourier analysis.

It will be convenient to make a few definitions. Let us call quadruples of the form $(x, x+a, x+b, x+a+b)$ *additive quadruples*. As commented above, these are in one-to-one correspondence with quadruples (x, y, z, w) such that $x+y = z+w$. They are also precisely the same as quadruples (x, y, z, w) such that $x - y = z - w$. We shall also call a triple of the form $(a, a + d, a + 2d)$ a *3-AP*. Let us define the *additive quadruple density* of a set A to be the quantity

$$\mathbb{E}_{x,a,b} A(x)A(x + a)A(x + b)A(x + a + b),$$

and the *3-AP density* to be the quantity

$$P_3(A) = \mathbb{E}_{x,d} A(x)A(x + d)A(x + 2d).$$

These are the probabilities that a random additive quadruple and a random 3-AP live entirely in A.

Lemma 6.6. *Let n be odd and let A be a subset of $\mathbb{Z}_n$ of density α. Suppose that $\max_{r\neq 0} |\hat{A}(r)| \leq c\alpha^2$. Then $|P_3(A) - \alpha^3| \leq c\alpha^3$.*

Proof. The 3-AP density is a quantity that has a nice expression on the Fourier side. To see this, first define a set A_2 to be the set of all u such that $u/2 \in A$. (It is here that we need n to be odd.) Note that

$$\hat{A}_2(r) = \mathbb{E}_x A_2(x)\omega^{-rx} = \mathbb{E}_x A(x/2)\omega^{-rx} = \mathbb{E}_x A(x)\omega^{-2rx} = \hat{A}(2r).$$

Indeed,

$$\begin{aligned} P_3(A) &= \mathbb{E}_{x+y=2z}A(x)A(y)A(z) \\ &= \mathbb{E}_{x+y=z}A(x)A(y)A(z/2) \\ &= \langle A*A, A_2\rangle \\ &= \langle \hat{A}^2, \hat{A}\rangle \\ &= \sum_r \hat{A}(r)^2\hat{A}(-2r). \end{aligned}$$

Since $\hat{A}(0) = \alpha$, it follows that

$$|P_3(A) - \alpha^3| = |\sum_{r\neq 0}\hat{A}(r)^2\hat{A}(-2r)| \leq \max_{r\neq 0}|\hat{A}(r)|\sum_{r\neq 0}|\hat{A}(r)||\hat{A}(-2r)|.$$

By Cauchy-Schwarz, the last sum is at most

$$(\sum_{r\neq 0}|\hat{A}(r)|^2)^{1/2}(\sum_{r\neq 0}|\hat{A}(-2r)|^2)^{1/2},$$

but the second of these factors is equal to the first, so we get $\sum_{r\neq 0}|\hat{A}(r)|^2$, which by Lemma 6.5 is $\alpha - \alpha^2 \leq \alpha$. The result follows. □

7. Roth's theorem

In this section we prove a famous result, the first non-trivial case of the Erdős-Turán conjecture, which was later to become Szemerédi's theorem. It is due to Klaus Roth.

Theorem 7.1. *For every $\delta > 0$ there exists n such that every subset $A \subset \{1,\dots,n\}$ of density at least δn contains an arithmetic progression of length 3.*

The basic structure of the proof is as follows. We begin by thinking of A as a subset not of $\{1,2,\dots,n\}$ but of $\mathbb{Z}_n$. Assuming that n is odd, we then use Lemma 6.6 to argue that either A contains an arithmetic progression of length 3 or there exists $r \neq 0$ such that $|\hat{A}(r)| \geq \delta^2/2$. In the second case, we find that A has a correlation with one of the functions ω_r with $r \neq 0$, and this "bias" enables us to deduce that the intersection of A with some arithmetic progression P of length around $\sqrt{n}$ has density at least $\delta + c\delta^2$. Having proved that, we can simply iterate the argument until either at some point the first case occurs or the density becomes so large in some subprogression that a simple averaging argument yields an arithmetic progression of length 3. (Or we can simply get to the point where the

first case *must* occur since otherwise there would be a subprogression inside which A had density greater than 1, a contradiction.)

Getting the details to work is slightly tricky, but this should not obscure the naturalness of the underlying ideas. We begin with a slight generalization of Lemma 6.6, which we need in order to cope with the fact that a 3-AP inside $\mathbb{Z}_n$ does not always correspond to a 3-AP inside $\{1, 2, \dots, n\}$.

Lemma 7.2. *Let n be odd and let A, B and C be subsets of $\mathbb{Z}_n$ with densities α, β and γ, respectively. Let $\max_{r\neq 0} |\hat{A}(r)| \leq \theta$. Then*

$$|\mathbb{E}_{x,d} A(x)B(x+d)C(x+2d) - \alpha\beta\gamma| \leq \theta(\beta\gamma)^{1/2}.$$

Proof. The proof is very similar to that of Lemma 6.6 so we will be brief about it. Following the earlier proof but changing some As to Bs and Cs, we can show that

$$\mathbb{E}_{x,d} A(x)B(x+d)C(x+2d) = \sum_r \hat{A}(r)\hat{B}(-2r)\hat{C}(r).$$

Then, again imitating the previous proof, we have that

$$|\mathbb{E}_{x,d} A(x)B(x+d)C(x+2d) - \alpha\beta\gamma| = |\sum_{r\neq 0} \hat{A}(r)\hat{B}(-2r)\hat{C}(r)| \leq \max_{r\neq 0} |\hat{A}(r)|(\beta\gamma)^{1/2}$$

and we are done. □

Corollary 7.3. *Let n be odd and let A be a subset of $\{1, 2, \dots, n\}$ of density δ. Suppose that $A \cap [n/3, 2n/3]$ has cardinality at least $\delta n/5$. Then either A contains an arithmetic progression of length 3, or when A is considered as a subset of $\mathbb{Z}_n$, there exists $r \neq 0$ such that $|\hat{A}(r)| \geq \delta^2/10$.*

Proof. Let $B = C = A \cap [n/3, 2n/3)$ and consider all of A, B and C as subsets of $\mathbb{Z}_n$. Note that any triple $(x, x+d, x+2d) \in A \times B \times C$ must have d belonging to the interval $(-n/3, n/3)$, from which it follows that $(x, x+d, x+2d)$ must be a 3-AP even when it is considered as a subset of $\{1, 2, \dots, n\}$. Therefore, if A contains no 3-AP, we obtain from Lemma 7.2 that $\theta(\beta\gamma)^{1/2} \geq \alpha\beta\gamma/2$, so $\theta \geq \alpha(\beta\gamma)^{1/2}/2$. By hypothesis, β and γ are at least $\delta/5$, and $\alpha = \delta$, so $\alpha(\beta\gamma)^{1/2}/2 \geq \delta^2/10$. □

The next step is to prove that if $|\hat{A}(r)| \geq \theta$ for some $r \neq 0$, then A has a slightly greater intersection than it should have with an arithmetic progression of length around $\sqrt{n}$. The key lemma we need is the following.

Lemma 7.4. *Let n be a positive integer, let $r \in \mathbb{Z}_n$, and let $\epsilon > 0$. Then there is a partition of $\{1, 2, \dots, n\}$ into arithmetic progressions P_i of length at least $\delta\sqrt{n}/16$ such that for every i and every $x, y \in P_i$ we have $|\omega^{rx} - \omega^{ry}| < \delta$.*

Proof. Let $m = \lfloor \sqrt{n} \rfloor$. Since the circumference of the unit circle is 2π, by the pigeonhole principle there exist distinct integers $u, v \in \{1, \dots, m\}$ such that $|\omega^{ru} - \omega^{rv}| \leq 2\pi m^{-1}$. Setting $d = |u - v|$, we then have that $|\omega^{r(x+d)} - \omega^{rx}| \leq 2\pi m^{-1}$ for every x. It follows that if P is any arithmetic progression of length t and common difference d, then $|\omega^{rx} - \omega^{ry}| \leq 2\pi t m^{-1}$ for every $x, y \in P$.

Now let us partition $\{1, 2, \dots, n\}$ into residue classes mod d. Then since $d \leq m \leq \sqrt{n}$, we can partition each residue class into arithmetic progressions of length between $\delta m/4\pi$ and $\delta m/2\pi$, and for each such progression P we have that the diameter of $\omega_r(P)$ is at most δ. □

Corollary 7.5. *Let n be a positive integer and let $A \subset \mathbb{Z}_n$ be a subset of density α. Suppose that there exists $r \neq 0$ such that $|\hat{A}(r)| \geq \theta$. Then there exists an arithmetic progression $P \subset \{1, 2, \dots, n\}$ of cardinality at least $\theta\sqrt{n}/32$ such that if we consider P as a subset of $\mathbb{Z}_n$, then $|A \cap P|/|P| \geq \alpha + \theta/4$.*

Proof. For each x, let $f(x) = A(x) - \alpha$. Then $\mathbb{E}_x f(x) = 0$ and $\hat{f}(r) = \hat{A}(r)$ for every $r \neq 0$. Therefore, by hypothesis there exists $r \in \mathbb{Z}_n$ such that $|\hat{f}(r)| \geq \theta$.

By Lemma 7.4 we can partition $\{1, 2, \dots, n\}$ into arithmetic progressions $P_1, \dots, P_r$, all of length at least $\theta\sqrt{n}/32$, such that the diameter of $\omega_r(P_i)$ is at most $\theta/2$ for each i.

We know that $|\mathbb{E}_x f(x)\omega^{-rx}| \geq \theta$, from which it follows that

$$\sum_i |P_i||\mathbb{E}_{x \in P_i} f(x)\omega^{-rx}| \geq \theta n = \theta \sum_i |P_i|.$$

For each i, let x_i be an element of P_i. Since the diameter of $\omega_r(P_i)$ is at most $\theta/2$ for each i, and $|f(x)| \leq 1$ for every x, we have for each i that

$$\begin{aligned} |\mathbb{E}_{x \in P_i} f(x)\omega^{-rx}| &\leq |\mathbb{E}_{x \in P_i} f(x)\omega^{-rx_i}| + \mathbb{E}_{x \in P_i}|f(x)||\omega^{-rx} - \omega^{-rx_i}| \\ &\leq |\mathbb{E}_{x \in P_i} f(x)| + \theta/2. \end{aligned}$$

Therefore,

$$\sum_i |P_i||\mathbb{E}_{x \in P_i} f(x)| \geq \frac{\theta}{2} \sum_i |P_i|.$$

We also have

$$\sum_i |P_i| \mathbb{E}_{x \in P_i} f(x) = \sum_x f(x) = 0.$$

It follows that there exists i such that

$$|P_i|(|\mathbb{E}_{x \in P_i} f(x)| + \mathbb{E}_{x \in P_i} f(x)) \geq \theta |P_i|/2.$$

For such an i, $\mathbb{E}_{x \in P_i} f(x)$ must be positive (or the left-hand side would be zero), so we find that

$$2|P_i| \mathbb{E}_{x \in P_i} f(x) \geq \theta |P_i|/2,$$

which implies that $\mathbb{E}_{x \in P_i} f(x) \geq \theta/4$. By the definition of f, this implies that

$$|A \cap P_i|/|P_i| \geq \alpha + \theta/4,$$

as we wanted. □

We have essentially finished the proof, but it remains to put all the ingredients together.

Lemma 7.6. *Let $A \subset \{1, 2, \dots, n\}$ be a set of density α and suppose that A does not contain an arithmetic progression of length 3. Then there is a subprogression $P \subset \{1, 2, \dots, n\}$ of cardinality at least $\alpha^2 \sqrt{n}/500$ such that $|A \cap P|/|P| \geq \alpha + \alpha^2/40$.*

Proof. If n is even, then partition $\{1, 2, \dots, n\}$ into the sets $\{1, 2, \dots, n/2-1\}$ and $\{n/2, n/2+1, \dots, n\}$, both of which have odd cardinality. In at least one of these, A has density at least α. Let $t = n$ if n is odd, and whichever of $n/2 - 1$ and $n/2 + 1$ is the length of the interval in which A has density at least α if n is even.

By translating if necessary, we may assume that A is a subset of $\{1, 2, \dots, t\}$. Corollary 7.3 now tells us that either A contains an arithmetic progression of length 3, or $A \cap [t/3, 2t/3)$ has cardinality less than $\alpha t/5$, or there exists $r \neq 0$ such that $|\hat{A}(r)| \geq \alpha^2/10$ when A is considered as a subset of $\mathbb{Z}_t$.

In the first case we are done. In the second, one of the sets $A \cap [1, t/3)$ and $A \cap [2t/3, t]$ has cardinality at least $2\alpha t/5$ and hence density at least $11\alpha t/10$. (The natural bound is $6\alpha t/5$ but this doesn't quite work because if t is a multiple of 3, then the set $A \cap [2t/3, t]$ has cardinality $t/3 + 1$ rather than $t/3$. But this is not at all important.) So in this case we are done by a long way.

In the third case, Corollary 7.3 gives us the hypothesis we need for Corollary 7.5 with $\theta = \alpha^2/10$. This then yields a progression P that satisfies the conclusion of the lemma, since $\theta \sqrt{t}/32 \geq \alpha^2 \sqrt{n}/500$. □

The above lemma represents one step of our iteration argument. Now let us finish the argument.

Proof of Theorem 7.1. As long as n is large enough, $\alpha^2\sqrt{n}/500 \geq n^{1/3}$. (More precisely, we need n to be at least $(500/\alpha^2)^6$.) Therefore, by Lemma 7.6, either A contains an arithmetic progression of length 3 or there is a progression P of size at least $n^{1/3}$ inside which A has relative density at least $\alpha + \alpha^2/40$. In the latter case, we simply repeat the argument.

If we have to repeat the argument $40/\alpha$ times, then the density will reach at least 2α. Then after a further $40/2\alpha$ iterations, it will reach 4α, and so on. So the total number of iterations we can need is at most $(40/\alpha)(1 + 1/2 + 1/4 + \dots) = 80/\alpha$.

From this it follows that A contains an arithmetic progression provided that after taking cube roots $80/\alpha$ times we still have a number greater than $(500/\alpha^2)^6$. That is, we need $n^{3^{-80/\alpha}} \geq (500/\alpha^2)^6$. Taking logs, this becomes $3^{-80/\alpha}\log n \geq 6(\log 500 + 2\log(\alpha^{-1})$. And taking logs again we find that the inequality

$$\log\log n \geq (80/\alpha)\log 3 + \log 6 + \log(\log 500 + 2\log(\alpha^{-1}))$$

suffices. This is at most C/α for some absolute constant C, so we have ended up showing that a density of $C/\log\log n$ suffices to guarantee a progression of length 3. This proves Roth's theorem. □

8. Szemerédi's regularity lemma

In this section I shall state and prove one of the most influential results in combinatorics: the famous regularity lemma of Szemerédi. His original version of the lemma, which was slightly different from the one people use today, formed an essential part of his proof of the Erdős-Turán conjecture, but has subsequently found a huge number of other applications. The underlying reason for this is quite simple: it provides a crude (but not too crude to be useful) classification of all graphs.

If one were to summarize the theorem in a few words, it would be to say that every graph can be approximately decomposed into a small number of quasirandom graphs. Soon I shall make that statement precise, and then I'll set about proving it. In a subsequent section I shall give a couple of applications. From the point of view of this course, it is really the applications that I am interested in, since the regularity lemma gives rise to what one might call the regularity method, which can be used to prove a number of results that are difficult to prove any other way. However, the proof of the lemma itself also introduces a useful

addition to the combinatorialist's toolbox: the notion of an energy-increment argument, which I shall explain later in this section.

I am now going to introduce a definition that is not quite standard, but I find it more convenient than the standard one. Let G be a bipartite graph with finite vertex sets X and Y. As earlier in these notes, if $A \subset X$ and $B \subset Y$, I shall write $G(A,B)$ for the induced bipartite subgraph with vertex sets A and B. I shall also write $d(A,B)$ for the density of this induced subgraph. It will also be useful to write $d(A)$ and $d(B)$ for the densities $|A|/|X|$ and $|B|/|Y|$ of A and B inside X and Y, respectively.

Definition 8.1. *Let G be a bipartite graph of density α with finite vertex sets X and Y and let $\epsilon > 0$. We say that G is* ϵ-regular *if for any two subsets $A \subset X$ and $B \subset Y$ we have the estimate*

$$|\mathbb{E}_{x\in X,y\in Y}G(x,y)A(x)B(y) - \alpha d(A)d(B)| \leq \epsilon.$$

This definition is closely related to Theorem 5.3. Indeed, suppose that G has 4-cycle density at most $\alpha^4(1+c^4)$. Then we find from that theorem that

$$|\mathbb{E}_{x\in X,y\in Y}G(x,y)A(x)B(y) - \alpha d(A)d(B)| \leq 2c\alpha(d(A)d(B))^{1/2},$$

which we can bound above very crudely by $2c\alpha$. So Theorem 5.3 gives us a sufficient condition for a bipartite graph to be ϵ-regular.

The definition can be adapted to graphs (as opposed to bipartite graphs) in the usual way: we simply take A and B to be subsets of the vertex set of G rather than one subset of each of the two vertex sets of G.

I feel a moral obligation to present the standard definition as well and prove that it is approximately equivalent to the one I have given. Returning to the bipartite case, it is usual to define the density $d(A,B)$ to be $\mathbb{E}_{x\in A,y\in B}G(x,y)$, and to say that G is ϵ-regular if $|d(A,B)-\alpha| < \epsilon$ whenever $d(A)$ and $d(B)$ are both at least ϵ.

Lemma 8.2. *The two definitions of ϵ-regularity are polynomially equivalent.*

Proof. Note first that

$$\mathbb{E}_{x\in X,y\in Y}G(x,y)A(x)B(y) = d(A)d(B)\mathbb{E}_{x\in A,y\in B}G(x,y) = d(A)d(B)d(A,B).$$

Therefore, if

$$|\mathbb{E}_{x\in X,y\in Y}G(x,y)A(x)B(y) - \alpha d(A)d(B)| \leq \epsilon,$$

it follows that

$$|d(A,B)-\alpha| \leq \epsilon/d(A)d(B).$$

Therefore, if $d(A)$ and $d(B)$ are both at least $\epsilon^{1/3}$ we find that $|d(A,B)-\alpha| \leq \epsilon^{1/3}$. Therefore, a graph that is ϵ-regular in the first sense is $\epsilon^{1/3}$ regular in the second sense.

In the reverse direction, if $d(A) < \epsilon$ or $d(B) < \epsilon$, then $|\mathbb{E}_{x\in X, y\in Y}G(x,y)A(x)B(y) - \alpha d(A)d(B)| < \epsilon$. If both $d(A)$ and $d(B)$ are at least ϵ and $|d(A,B)-\alpha| \leq \epsilon$, then $|d(A)d(B)d(A,B) - \alpha d(A)d(B)| \leq \epsilon d(A)d(B) \leq \epsilon$. So a graph that is ϵ-regular in the second sense is ϵ-regular in the first sense. □

I prefer the first definition because it is better suited to analytic proofs of the kind I am stressing in this course. However, the first is also convenient in many situations, and since it is standard, it is important to know it.

The next definition we need is that of an ϵ-regular *partition*, or rather pair of partitions. If G is a bipartite graph with finite vertex sets X and Y, and $X_1,\dots,X_r$ and $Y_1,\dots,Y_s$ are partitions of X and Y, then we call the partitions *ϵ-regular* if

$$\sum\{d(X_i)d(Y_j) : G(X_i,Y_j) \text{ is not } \epsilon\text{−regular}\} \leq \epsilon.$$

A helpful way of thinking about this is as follows. If $A \subset X$ and $B \subset Y$, then say that (A,B) is an *ϵ-regular pair* if the graph $G(A,B)$ is ϵ-regular. Then the partitions above are ϵ-regular if the probability that a random edge belongs to a pair of cells that form an ϵ-regular pair is at least $1-\epsilon$.

We are now ready to state Szemerédi's regularity lemma for bipartite graphs.

Theorem 8.3. *Let G be a bipartite graph with finite vertex sets X and Y and let $\epsilon > 0$. Then there exists an ϵ-regular pair of partitions $X_1,\dots,X_r$ of X and $Y_1,\dots,Y_s$ of Y such that $r,s \leq K(\epsilon)$.*

To be completed.

9. The triangle removal lemma and the corners theorem

To be written.

10. Bohr sets and Bogolyubov's method

Lemma 10.1. *Let $r_1,\dots,r_k \in \mathbb{Z}_n$ and let $0 < \delta \leq 2$. Let θ be the proportion of points z in the unit circle such that $\Im(z) \geq 0$ and $|1-z| \leq \delta$. Then the Bohr set $B(r_1,\dots,r_k;\delta)$ has density at least θ^k.*

Proof. For each $\phi \in [0,1)$ let $A_\theta(\phi)$ be the arc of the unit circle that goes from $e^{2\pi i\phi}$ to $e^{2\pi i(\phi+\theta)}$. If we choose $(\phi_1,\dots,\phi_k)$ uniformly at random from $[0,1)^k$, then for each x the

probability that $(\omega^{r_1x},\dots,\omega^{r_kx}) \in A_\theta(\phi_1)\times\cdots\times A_\theta(\phi_k)$ is θ^k. It follows that there exists $(\phi_1,\dots,\phi_k)$ such that the density of x such that $(r_1x,\dots,r_kx) \in A_\theta(\phi_1)\times\cdots\times A_\theta(\phi_k)$ is at least θ^k.

If x and y both have that property, then for each i we have that $\omega^{r_i(x-y)}$ belongs to the arc that goes from $e^{-2\pi i\theta}$ to $e^{2\pi i\theta}$. It follows that $|1-\omega^{r_i(x-y)}| \leq \delta$ for each i, and therefore that $x-y \in B(r_1,\dots,r_k;\delta)$. The result follows. □

Note that θ is given by the formula $\delta = 2\sin(\pi\theta)$. In particular, $\theta > \delta/2\pi$. That will be useful in the next lemma. Note also that $B(r_1,\dots,r_k;\delta)$ is equal to the set of all x such that $r_ix \in [-\theta n,\theta n]$ for each i.

Corollary 10.2. *Let n be prime and let $r_1,\dots,r_k \in \mathbb{Z}_n$ and $\delta > 0$. Then the Bohr set $B(r_1,\dots,r_k;\delta)$ contains an arithmetic progression mod n of length at least $\delta n^{1/k}/2\pi$.*

Proof. We begin by finding a small ρ such that $B(r_1,\dots,r_k;\rho)$ contains a non-zero point. For this it is sufficient if it has cardinality greater than 1, and by Lemma 10.1 and the remark immediately following it, that will happen if $(\rho/2\pi)^k \geq n^{-1}$, which is true if $\rho = 2\pi n^{-1/k}$.

If x belongs to the Bohr set $B(r_1,\dots,r_k;\rho)$ and $m\rho \leq \delta$, then the arithmetic progression $\{0,x,2x,\dots,mx\}$ is a subset of $B(r_1,\dots,r_k;\delta)$. This proves the result. □

For the next lemma we need a couple more definitions. A d-dimensional *lattice* is a subgroup of $\mathbb{R}^d$ that is spanned by n linearly independent points. (Equivalently, it is a discrete subgroup of $\mathbb{R}^d$ that is not contained in a proper subspace.) A d-dimensional *lattice convex body* is the intersection of a d-dimensional lattice with a convex set. It is *symmetric* if it is invariant under the transformation $x \mapsto -x$.

Lemma 10.3. *Let $r_1,\dots,r_k$ be non-zero elements of $\mathbb{Z}_n$ and let $0 < \delta < \sqrt{2}$. Then the Bohr set $B(r_1,\dots,r_k;\delta)$ is Freiman isomorphic to a k-dimensional symmetric lattice convex body.*

Proof. Let Λ be the lattice generated by $n\mathbb{Z}^k$ together with the point $(r_1,\dots,r_k)$. That is, Λ consists of all points $(u_1,\dots,u_k)$ that are congruent mod n to a point of the form $(r_1x,\dots,r_kx)$.

Now let K be the convex body $[-\theta n,\theta n]^k$, where θ is as in Lemma 10.1. Note that $\theta < 1/4$. We shall show that the lattice convex body $\Lambda \cap K$ is Freiman isomorphic to $B(r_1,\dots,r_k;\delta)$.

The isomorphism is simple to define. Given $x \in \mathbb{Z}$, write $R(x)$ for its least residue mod n, meaning the residue with smallest modulus (taking $n/2$ if it happens that $x \equiv n/2$). Then we map $x \in B(r_1, \dots, r_k; \delta)$ to $(R(r_1x), \dots, R(r_kx))$. This belongs to Λ and also to K.

The map just defined can be inverted as follows. Given a point $(u_1, \dots, u_k) \in \Lambda \cap K$, we know that $(u_1, \dots, u_k)$ is congruent mod n to a point of the form $(r_1x, \dots, r_kx)$. Therefore, the only possibility for x is $r_1^{-1}u_1$ (which is congruent mod n to $r_i^{-1}u_i$ for each i). But then for each i we find that $u_i = R(r_ix)$, and therefore that $R(r_ix) \in [-\theta n, \theta n]$, which implies that $|1 - \omega^{r_ix}| \leq \delta$.

This inverse map is the restriction of a group homomorphism, so it is certainly a Freiman homomorphism. It remains to prove that the original map is a Freiman homomorphism. But suppose that $x, y, z, w \in B(r_1, \dots, r_k; \delta)$ and that $x + y = z + w$. Then for each i we have that

$$R(r_i(x + y)) = R(r_i(z + w)).$$

Since r_ix and r_iy live in $[-\theta n, \theta n]$ and $\theta < 1/4$, it follows that $R(r_i(x+y)) = R(r_ix)+R(r_iy)$, and similarly for z and w. Therefore, for each i we have

$$R(r_ix) + R(r_iy) = R(r_iz) + R(r_iw).$$

It follows that the original map is indeed a Freiman homomorphism. □

For the next lemma it will be convenient to write $B[K; \theta]$ for the set of all x such that $rx \in [-\theta n, \theta n]$ for every $r \in K$.

Lemma 10.4. *Let K be a subset of $\mathbb{Z}_n$ of size k and let $\theta > 0$. Then $|B[K, 2\theta]| \leq 4^k|B[K, \theta]$.*

Proof. Cover the interval $[-\theta n, \theta n]$ by four intervals $I_1, \dots, I_4$ of width $\theta n/2$. Let $K = \{r_1, \dots, r_k\}$, and for each $h = (h_1, \dots, h_k) \in \{1, 2, 3, 4\}^k$, let $B(h)$ be the set of x such that $r_jx \in I_{h_j}$ for every j. Then the union of the 4^k sets $B(h)$ is equal to $B[K; 2\theta]$ and $B(h) - B(h) \subset B[K; \theta]$ for each h. The result follows. □

Our next aim is to prove a very useful fact about Bohr sets, which roughly speaking says that for many θ the size of $B[K, \theta]$ has good continuity properties. Before we give the statement, we prove *Vitali's covering lemma*.

Lemma 10.5. *Let $[a, b]$ be a real interval and let $\mathcal{U}$ be an open covering of $[a, b]$. Then there are disjoint sets $U_1, \dots, U_k$ in $\mathcal{U}$ that cover at least half of $[a, b]$.*

Proof. Let $\mathcal{V}$ be a minimal subcover. Since $[a,b]$ is compact, $\mathcal{V}$ is finite. Also, the minimality of $\mathcal{V}$ implies that for each $U \in \mathcal{V}$ there exists $x \in [a,b]$ such that $x \in U$ and x belongs to no other set in V. Let $\mathcal{V} = \{U_1, \dots, U_k\}$ and for each i let $x_i \in U_i \setminus \bigcup_{j \neq i} U_j$. Without loss of generality $x_1 < x_2 < \cdots < x_k$.

For each $1 \leq i < j \leq k$ we have that $x_i < x_j$, $x_i \in U_i$, $x_j \notin U_i$, $x_j \in U_j$, $x_j \notin U_i$. It follows that all points in U_i are less than x_j and all points in U_j are greater than x_i. From this it follows that the sets $U_1, U_3, U_5, \dots$ are disjoint, as are the sets $U_2, U_4, U_6, \dots$. But one or other of these collections of sets must cover at least half of $[a,b]$. □

Let f be a function from an interval $[a,b]$ to $\mathbb{R}$, let $x \in [a,b]$ and let $C \geq 0$ be a constant. We shall say that f is *C-Lipschitz at x* if $|f(y) - f(x)| \leq C|y - x|$ for every $y \in [a,b]$.

Corollary 10.6. *Let $f : [a,b] \to \mathbb{R}$ be an increasing function with $f(b) - f(a) = \lambda(b-a)$. Let $a_1 = (3a+b)/4$ and let $b_1 = (a+3b)/4$. Then there exists $x \in [a_1, b_1]$ such that f is 4λ-Lipschitz at x.*

Proof. Suppose that this is not true. Then for every $x \in [a_1, b_1]$ we can find $y \in [a,b]$ such that $|f(y) - f(x)| > 4\lambda|y - x|$. Since f is increasing, this allows us to find an open interval (u_x, v_x) containing x such that $f(v_x) - f(u_x) > 4\lambda(v_x - u_x)$. By Lemma 10.5 we can cover at least half of $[a_1, b_1]$ with a disjoint collection of such intervals, from which it follows that $f(b_1) - f(a_1) > 2\lambda(b_1 - a_1) = \lambda(b-a)$. From that it follows that $f(b) - f(a) > \lambda(b-a)$, a contradiction. □

Corollary 10.7. *Let $K \subset \mathbb{Z}_n$ be a set of size k. Then for every $\rho > 0$ there exists $\alpha \in [1/4, 3/4]$ such that the function $f : [0,1] \to \mathbb{R}$ defined by $f : x \mapsto \log_2 |B[K; 2^x \rho]|$ is $8k$-Lipschitz at α.*

Proof. Clearly f is an increasing function. Also, by Lemma 10.4, $f(1) - f(0) = \log_2 |B(K; 2\rho]| - \log_2 |B[K;\rho]| \leq 2k$. Corollary 10.6 then gives us the required α. □

Now let us translate Corollary 10.7 into a statement about the size of Bohr sets. Let K, ρ and α be as in the corollary, and let $\theta = 2^\alpha \rho$. Then for every $\beta > 1$ we have the inequalities

$$|B[K; \beta\theta]| \leq 2^{8k\beta} |B[K;\theta]|$$

and

$$|B[K; \beta^{-1}\theta] \geq 2^{-8k\beta} |B[K;\theta]|.$$

For small $\delta > 0$, we can deduce from this that

$$|B[K; (1+\delta)\theta]| \leq (1 + 8k\delta) |B[K;\theta]|$$

and

$$|B[K;(1-\delta)\theta] \geq (1-8k\delta)|B[K;\theta]|.$$

11. Plunnecke's theorem and Ruzsa's embedding lemma

Lemma 11.1. *Let A and B be finite subsets of an Abelian group G and suppose that $|A+B| = K|A|$ and that $|Z+B| \geq K|Z|$ for every $Z \subset A$. Then $|A+B+C| \leq K|A+C|$ for every finite set $C \subset G$.*

Proof. We shall prove this by induction on C. If C is a singleton, then $|A+B+C| = |A+B|$ and $|A+C| = |A|$ so we are done by our hypothesis.

Suppose now that we have the result for C and let us try to prove it for $C' = C \cup \{x\}$ for some arbitrary element $x \in G$. We have $A+B+C' = (A+B+C) \cup (A+B+x)$. But we can say more than this. Let W be the set of $a \in A$ such that $a+x \in A+C$. Then $W+x \subset A+C$, so $W+B+x \subset A+B+C$. It follows that

$$A+B+C' = (A+B+C) \cup ((A+B+x) \setminus (W+B+x)).$$

By induction, we have that $|A+B+C| \leq K|A+C|$. We also have that $|A+B+x| = K|A|$ and that $|W+B+x| \geq K|W|$. Therefore,

$$|A+B+C'| \leq K(|A+C| + |A| - |W|).$$

But $A+C' = (A+C) \cup ((A \setminus W)+x)$, and this is a disjoint union. Therefore,

$$|A+C'| \geq |A+C| + |A| - |W|,$$

which completes the proof. □

Corollary 11.2. *Let A and B be finite subsets of an Abelian group G and suppose that $|A+B| \leq C|A|$. Then $|hB| \leq C^h|A|$.*

Proof. Let $A' \subset A$ be such that the ratio $|A'+B|/|A'|$ is minimized. Then $|A'+B| \leq C|A|$ and A' and B satisfy the conditions of the previous lemma. It follows that $|A' + B + (h-1)B| \leq C|A' + (h-1)B|$ for every h. Therefore, by induction we have that $|A'+hB| \leq C^h|A'|$, which implies the result. In particular, if we set $A = B$, this tells us that if $|A+A| \leq C|A|$, then $|hA| \leq C^h|A|$. □

We now want to get a similar result for sums and differences. For that we can use a very useful lemma of Ruzsa, known as the *Ruzsa triangle inequality*.

Lemma 11.3. *Let A, B and C be finite subsets of an Abelian group. Then $|A-B||B-C| \geq |B||A-C|$.*

The reason this is called a triangle inequality is that we can define a sort of "distance" between two sets A and B to be $d(A,B) = |A-B|/|A|^{1/2}|B|^{1/2}$. Then we can reformulate as saying that $d(A,C) \leq d(A,B)d(B,C)$. (Thus the inequality is multiplicative rather than additive. It should be stressed that even after taking logs we do not actually get a metric, since $d(A,A)$ is 1 only if A is a coset of a subgroup.)

Proof. We shall prove this result by defining an injection $\phi : B \times (A-C) \to (A-B) \times (B-C)$. This we do in a fairly obvious way. First, for each $u \in A-C$ we choose elements $\rho(u)$ of A and $\sigma(u)$ of C with $\rho(u) - \sigma(u) = u$. Then we define $\phi(b,u)$ to be $(\rho(u) - b, b - \sigma(u))$.

Suppose now that we are presented with a pair (r,s) and told that it is $\phi(b,u)$ for some $(b,u) \in B \times (A-C)$. Then $r+s = \rho(u) - \sigma(u) = u$. But once we have determined u, we also determine b, since it equals $s + \sigma(u)$. □

Corollary 11.4. *Let A be a finite subset of an Abelian group and suppose that $|A+A| \leq C|A|$ or that $|A-A| \leq C|A|$. Then for every r and s we have that $|rA - sA| \leq C^{r+s}|A|$.*

Proof. If $|A+A| \leq C|A|$, then by the proof of Lemma 11.2 we know that A has a subset A' such that $|A' + rA| \leq C^r|A'|$ and $|A' + sA| \leq C^s|A'|$. Therefore, by Lemma 11.3 with the roles of A, B and C played by rA, $-A'$ and sA, respectively (noting that $|A' + sA| = |-A' - sA|$), we find that $|A' + rA||A' + sA| \geq |A'||rA - sA|$, which implies that $|rA - sA| \leq C^{r+s}|A'|$, which is at most $C^{r+s}|A|$. □

Lemma 11.5. *Let A and B be sets such that $|A+B| \leq C|B|$. Then there is a set X of size at most C such that $A \subset X + B - B$.*

Proof. Let $X = \{x_1, \dots, x_k\}$ be a maximal subset of A such that the sets $x_i + B$ are disjoint. Since each set $x_i + B$ has size $|B|$ and is a subset of $A + B$, X has size at most C. The maximality of X tells us that for every $a \in A$ there exists i such that $a + B$ and $x_i + B$ are not disjoint, which is the same as saying that $a \in x_i + B - B$. The result follows. □

12. The polynomial method

It is not very easy to say exactly what the polynomial method is, but in broad terms it is exploiting facts about zeros of polynomials to prove results that do not appear to have anything to do with polynomials. Typically, one tries to prove that a small combinatorial structure cannot exist by showing that if it did, then it would allow us to define a non-zero

polynomial of low degree that vanishes on a large set, sometimes with some extra structure, which we can then contradict by proving results that show that non-zero polynomials of low degree cannot vanish on such sets.

In this section we shall illustrate the method with three results. The first is a celebrated result of Dvir, who solved the so-called Kakeya problem for finite fields. This is the following question: suppose that A is a subset of $\mathbb{F}_p^n$ that contains a translate of every line. Must A have size $p^{n-o(1)}$? Here, we are taking n to be fixed and p to be tending to infinity. Dvir's solution gave the following theorem.

Theorem 12.1. *Let A be a subset of $\mathbb{F}_p^n$ that contains a translate of every line. Then $|A| \geq c(n)p^n$.*

The proof needs a couple of simple (but surprisingly powerful) lemmas.

Lemma 12.2. *Let $A \subset \mathbb{F}_p^n$ be a set of size less than $\binom{n+d}{d}$. Then there exists a non-zero polynomial $P(x_1, \dots, x_n)$ of degree d that vanishes on A.*

Proof. A polynomial of degree d in the variables $x_1, \dots, x_n$ is a linear combination of monomials of degree at most d. The number of monomials of degree k is the number of ways of writing a number less than or equal to d in the form $a_1 + \cdots + a_n$ with each a_i non-negative. By a standard holes-and-pegs argument, this is $\binom{n+d}{d}$. (Given $n+d$ holes with d pegs, then a_i is the number of pegs that follow the ith hole.) Therefore, for a polynomial of degree d to vanish at m given points, a certain set of m linear combinations of the coefficients must all be zero. By dimension considerations, this is possible if m is less than the number of coefficients. The result follows. □

When $d = p-1$, this gives us that for every set A of size less than $\binom{n+p-1}{p-1} = \binom{n+p-1}{n}$ we can find a non-zero polynomial of degree d that vanishes on A. Since $\binom{n+p-1}{p-1} > p^n/n!$, this is in particular true when $|A| \leq p^n/n!$.

The next lemma is the main idea behind the proof of Dvir's theorem.

Lemma 12.3. *Suppose that $A \subset \mathbb{F}_p^n$ contains a line in every direction, that $d < p$, and that there exists a non-zero polynomial f of degree at most d that vanishes on A. Then there is a non-zero degree-d polynomial that vanishes everywhere on $\mathbb{F}_p^n$.*

Proof. Without loss of generality the degree of f is exactly d. Let $a, z \in \mathbb{F}_p^n$ with $z \neq 0$ and let L be the line consisting of all points $a + tz$. The restriction of f to L is a polynomial of degree d in t, and its leading coefficient is $f_d(z)$, where f_d is the degree-d part of f. To

see this, observe that for any monomial $\prod_{i=1}^n x_i^{r_i}$ its value at $a+tz$ is $\prod_{i=1}^n (a_i+tz_i)^{r_i}$, so if the monomial has degree d, then to obtain a term in t^d we must choose tz_i from each bracket, which gives $t^d \prod_{i=1}^n z_i^{r_i}$.

Now if f vanishes everywhere on L, then since its dependence on t is given by a polynomial of degree less than p, all the coefficients of that polynomial must be zero. It follows that $f_d(z)=0$. But z was an arbitrary non-zero element of $\mathbb{F}_p^n$, and f_d vanishes at zero as well, so it vanishes everywhere. □

To finish the proof, we need the second of our simple but powerful lemmas. It is called the Schwartz-Zippel lemma, and it tells us that a polynomial of degree d in n variables cannot have too many roots. Some sort of intuition for how many roots we might expect comes from thinking about linear polynomials: there we do not get more than p^{n-1} roots. A very useful and general principle in applications of arithmetic geometry is that polynomials behave in a rather similar way to linear functions. For example, a linear function from $\mathbb{R}$ to $\mathbb{R}$ has at most one root, while a polynomial of degree d has at most d roots, which is the same up to a constant.

Before we prove the Schwartz-Zipple lemma, we need a preparatory result. Note that there is an important distinction between the zero polynomial, which is the polynomial with all coefficients of all monomials equal to zero, and a polynomial that takes the value zero everywhere on $\mathbb{F}_p^n$. For instance, the polynomial $x^{p-1}-1$ is not the zero polynomial but takes the value zero everywhere on $\mathbb{F}_p$.

Lemma 12.4. *Let f be a non-zero polynomial on $\mathbb{F}_p^n$ of degree less than p. Then f is not identically zero.*

Proof. We shall prove this by induction on n. If $n=1$, then a non-zero polynomial that vanishes everywhere has p roots, so must be of degree p. Essentially the same argument works for general n. If f vanishes everywhere, then for each a it vanishes on the set $x_1=a$. But the restriction of f to that set is a polynomial of degree less than p in the variables $x_2,\dots,x_n$, so by induction it is the zero polynomial.

In other words, if we substitute a for x_1 in the definition of f, we obtain the zero polynomial. If we think of f as a polynomial in x_1 with coefficients in $\mathbb{F}_p[x_2,\dots,x_n]$, then the polynomial division algorithm tells us that we can write $f(x)$ in the form $P(x_1,\dots,x_n)(x_1-a)+Q(x_2,\dots,x_n)$, so if this polynomial is the zero polynomial when we substitute $x_1=a$ we obtain that Q is the zero polynomial and (x_1-a) divides f. But if this is true for every a, then again we find that f has to have degree at least p, contradicting our assumption. □

Lemma 12.5. *Let f be a non-zero polynomial of degree at most d on $\mathbb{F}_p^n$. Then f has at most dp^{n-1} roots.*

Proof. Without loss of generality the degree is exactly d. As we have already seen, if we restrict f to the line consisting of points $a+tz$, then we obtain a polynomial of degree d in the single variable t with leading coefficient $f_d(z)$, where f_d is the degree-d part of f. By Lemma 12.4 we may choose z such that $f_d(z)\neq 0$. But that means that on every line L in direction z the restriction of f to L is given by a polynomial in one variable of degree d. So f has at most d roots in any line, and therefore at most dp^{n-1} roots in total. □

It follows from the Schwartz-Zippel lemma that a non-zero polynomial of degree less than p cannot vanish on all of $\mathbb{F}_p^n$. Combining this with Lemmas 12.2 and 12.3 we obtain Theorem 12.1 with $c(n)=1/n!$.

We now turn to a result due to Noga Alon, known as the combinatorial Nullstellensatz.

Lemma 12.6. *Let f be a non-zero polynomial in n variables over $\mathbb{F}_p$ of degree $k_1+\cdots+k_n$, where the k_i are non-negative integers and the coefficient of $x_1^{k_1}\dots x_n^{k_n}$ is non-zero. Let $S_1,\dots,S_n$ be subsets of $\mathbb{F}_p$ such that $|S_i|>k_i$ for each i. Then f does not vanish on $S_1\times\cdots\times S_n$.*

Proof. We prove this by induction on the degree of f. The result is easy when the degree is zero.

If the degree is greater than zero, then without loss of generality $k_1>0$. Let $a\in S_1$ and use polynomial division to write $f(x)$ in the form $(x_1-a)P(x)+Q(x)$, where Q does not depend on x_1. Since the term in $x_1^{k_1}\dots x_n^{k_n}$ has a non-zero coefficient in f, and the degree of P is $k_1+\cdots+k_n-1$, the term in $x_1^{k_1-1}x_2^{k_2}\dots x_n^{k_n}$ has non-zero coefficient in P.

Suppose that the result is false. Then f vanishes on $\{a\}\times S_2\times\cdots\times S_n$, from which it follows that Q vanishes on this set too. Since Q does not depend on x_1, it vanishes on all of $S_1\times\cdots\times S_n$. Therefore, $(x_1-a)P$ vanishes on $S_1\times\cdots\times S_n$, which implies that P vanishes on $(S_1\setminus\{a\})\times S_2\times\cdots\times S_n$. This contradicts the inductive hypothesis. □

As our first application of the combinatorial Nullstellensatz, we give a short proof of the Cauchy-Davenport theorem, which is the following result (discovered independently by Cauchy in 1813 and Davenport in 1935 – apparently it was not until 1947 that Davenport found out that Cauchy had beaten him to it by over a century). Neither Cauchy nor Davenport used the method below.

Theorem 12.7. *Let p be a prime and let A and B be subsets of $\mathbb{F}_p$. Then $|A+B| \geq \min\{p, |A|+|B|-1\}$.*

Proof. Once we have the clue that this can be proved using the combinatorial Nullstellensatz, we need to find a suitable sequence of sets $S_1, \dots, S_n$ and a polynomial that vanishes on $S_1 \times \cdots \times S_n$ and that has small degree unless $A+B$ is large.

This gives enough of a clue to complete the proof. The obvious sequence of sets to take, since the only sets we have around are A and B, is (A, B). We want the degree of our polynomial to depend on the size of $A+B$, and we also want it to vanish on $A \times B$. The most economical way of getting it to vanish at a point (a, b) is to ensure that the polynomial has a factor $x+y-(a+b)$, which leads to the idea of considering the polynomial

$$f(x,y) = \prod_{c \in A+B} (x+y-c).$$

This vanishes on $A \times B$ and has degree equal to $|A+B|$, so it looks promising.

Suppose now that $|A+B| < p$ and $|A+B| \leq |A|+|B|-2$. We want to contradict the combinatorial Nullstellensatz, so we need a monomial $x^r y^s$ with non-zero coefficient with $r < |A|$, $s < |B|$ and $r+s = |A+B|$. But if we pick any r and s that satisfy the last three conditions, which we clearly can if $|A+B| \leq |A|+|B|-2$, then the coefficient of $x^r y^s$ in the polynomial is $\binom{r+s}{r}$, and this is non-zero because p is prime and $r+s < p$. □

Note that this result is sharp if A and B are arithmetic progressions in $\mathbb{F}_p$ with the same common difference. Note too that the result is false in $\mathbb{Z}_n$ if n is composite, since then $\mathbb{Z}_n$ has proper subgroups

Now we shall use the combinatorial Nullstellensatz to prove a variant of the Cauchy-Davenport theorem. The result is due to da Silva and Hamidoune, who found a somewhat involved combinatorial proof. This short argument was discovered by Alon, Nathanson and Ruzsa. Let us write $A\dot{+}B$ for the set of sums $a+b$ such that $a \in A$, $b \in B$ and $a \neq b$.

Theorem 12.8. *Let p be a prime and let A and B be subsets of $\mathbb{Z}_p$. Then $|A\dot{+}B| \geq \min\{p, |A|+|B|-3\}$.*

Proof. First we show that if $|A|+|B| \geq p+2$, then $A\dot{+}B = \mathbb{Z}_p$. If $p = 2$ the result is trivial. To see it for $p > 2$, let $x \in \mathbb{Z}_p$. Then $A \cap (x-B)$ contains at least two elements, so at least one of them, a does not satisfy the equation $2a = x$. But then $a \in A$ and $b = x-a \in B$ are distinct elements that add up to x.

We would now like to apply the combinatorial Nullstellensatz, so we need to show that if $A\dot{+}B$ is too small, then some polynomial of low degree vanishes everywhere on a product

set. The natural product set to try to take is $A+B$. What low-degree polynomial would vanish on $A+B$? Well, we know that $A\dot{+}B$ is small, so a first approximation would be to take the polynomial $\prod_{c\in A\dot{+}B}(x+y-c)$. The trouble with that is that it does not necessarily vanish when $x=y\in A\cap B$. But we can take care of that by simply multiplying by the polynomial $x-y$.

So now we have a polynomial of degree $|A\dot{+}B|+1$ that vanishes on $A\times B$. For technical reasons it will be convenient to modify it slightly. Let us assume that the result is false and let C be a set that contains $A\dot{+}B$ and has cardinality exactly $|A|+|B|-4$. Let P be the polynomial $(x-y)\prod_{c\in C}(x+y-c)$. This polynomial vanishes on $A\dot{+}B$ and has degree $|A|+|B|-3$.

Let us look at the terms in $x^{|A|-1}y^{|B|-2}$ and $x^{|A|-2}y^{|B|-1}$, since if either of these has a non-zero coefficient then we will have contradicted the combinatorial Nullstellensatz. Let us write $r=|A|$, $s=|B|$, $t=r+s-3$. Then the coefficient of $x^{r-1}y^{s-2}$ is $\binom{t-1}{r-2}-\binom{t-1}{r-1}$ and the coefficient of $x^{r-2}y^{s-1}$ is $\binom{t-1}{r-3}-\binom{t-1}{r-2}$. It is not possible for three consecutive (or indeed any three) binomial coefficients to be equal, so the result is proved. □

Note that if $A=B=\{0,1,2,\dots,r-1\}$, then $A\dot{+}B=\{1,2,\dots,2r-3\}$, so the result is sharp when the sets have equal size. It is not sharp for general sizes, since for example if B is a singleton, then $|A|+|B|\geq|A|-1=|A|+|B|-2$.

The above two results can be proved by other means, but there are results that have been proved using the combinatorial Nullstellensatz for which no other proof is known – just as no proof of Dvir's theorem is known that does not go via polynomials.