SZEMERÉDI REGULARITY LEMMA AND ARITHMETIC PROGRESSIONS

ANNIE MARSDEN

Abstract. This paper consists of two main components. The first component presents the Szemerédi Regularity Lemma and uses it to give a proof of the Szemerédi theorem for arithmetic progressions of length 3. The second component explores the Szemerédi Regularity Lemma in the context of 3-uniform hypergraphs and then goes on to prove Szemerédi's theorem for arithmetic progressions of length 4.

Contents

1. Introduction

The Szemerédi Regularity Lemma came about as a lemma to prove the conjecture of Erdős and Turán that one can always find long arithmetic progressions in sequences of integers of positive upper density. Ultimately it has been one of the most important tools in graph theory. The main idea of the Regularity Lemma is that every graph can be approximated by a random graph. Given any graph we can partition the graph in such a way so that the partitions can be viewed as vertices and the density between the partitions can be viewed as the probability of an edge between the corresponding vertices. We can view the question of the existence of an arithmetic progression of length 3 as a question about the existence of a triangle in a graph. We use the Regularity Lemma to help us answer this question. In a similar manner we can view the question of the existence of an arithmetic progression of length 4 as a question about the existence of a simplex in a 3-uniform hypergraph. We use a hypergraph version of the Regularity Lemma to help us answer this question. By modifiying the definition of regularity and the Szemerédi Regularity Lemma to extend to k-uniform hypergraphs, this method works to prove the existence of an arithmetic progression of length k. In this paper we go over the proof for arithmetic progressions of length 3 and length 4.

Date: March 16, 2015.

2. Szemerédi Regularity Lemma

Definition 2.1. (Density) Given a graph G with vertex set V and two disjoint vertex sets $A \subset V$, $B \subset V$, the *density* of the (A, B)-pair, $d(A, B)$, is defined as follows:

$$d(A,B) = \frac{e(A,B)}{|A| \cdot |B|}$$

Where $e(A, B)$ is the number of edges between A and B, $|A|$ is the number of vertices in A, $|B|$ is the number of vertices in B.

Definition 2.2. (ϵ-Regularity) Let $\epsilon > 0$. Given a graph G with vertex set V and two disjoint vertex sets $A \subset V$, $B \subset V$, the (A, B)-pair is ϵ-regular if $\forall\ X \subset A$ and $\forall\ Y \subset B$ we have that

$$(|X| > \epsilon|A| \text{ and } |Y| > \epsilon|B|) \implies (|d(X,Y) - d(A,B)| < \epsilon)$$

.

Theorem 2.3. *(Szemerédi Regularity Lemma as stated in Komlós and Simonovits Paper) For every $\epsilon > 0$ and for every $m > 0$ there exist two integers $M(\epsilon, m)$ and $N(\epsilon, m)$ with the following property: for every graph G with $n \geq N(\epsilon, m)$ vertices there is a partition of the vertex set into $k + 1$ classes*

$$V = V_0 + V_1 + V_2 + ... + V_k$$

such that
(1) $m \leq k \leq M(\epsilon, m)$
(2) $|V_0| < \epsilon n$
(3) $|V_1| = |V_2| = ... = |V_k|$
(4) all but at most ϵk^2 of the pairs (V_i, V_j) are ϵ-regular.

Definition 2.4. Let G and H be graphs. A function $\phi : V(H) \to V(G)$ is an isomorphic embedding if $(u, v) \in E(G) \iff (\phi(u), \phi(v)) \in E(H)$

Theorem 2.5. *(Counting Lemma as stated in Gower's Paper) For every $\alpha > 0$ and every k there exists $\epsilon > 0$ with the following property. Let $V_1, ..., V_k$ be sets of vertices in a graph G, and suppose that for each pair (i,j) the pair (V_i, V_j) is ϵ-regular with density d_{ij}. Let H be a graph with vertex set $(x_1, ..., x_k)$, let $v_i \in V_i$ be chosen independently and uniformly at random, and let ϕ be the map that takes x_i to v_i for each i. Then the probability that ϕ is an isomorphic embedding differs from $\prod_{x_i x_j \in H} d_{ij} \prod_{x_i x_j \notin H} (1 - d_{ij})$ by at most α.*

3. Extension of Szemerédi Regularity Lemma to Arithmetic Progressions of Length Three

Lemma 3.1. *(Triangle Removal Lemma) For every constant $a > 0$ $\exists c > 0$ such that if G is any graph with n vertices that contains at most cn^3 triangles, then it suffices to remove an^2 edges from G to make it triangle free.*

Proof. Choose $\epsilon > 0$ and $m > 0$ such that $\epsilon + \frac{1}{m} \leq \frac{a}{2}$, the reason for this will be clear deeper in the proof. By the regularity lemma there exists an ϵ-regular partition $V_1, ..., V_k$, where $m \leq k$. From this partition we remove the following edges:

(1) All edges between V_i, V_j if V_i, V_j is not an ϵ-regular pair

(2) Edges within each V_i
(3) All edges between V_i, V_j if $|e(V_i, V_j) \leq a|V_i||V_j|$

Now let us determine an upper bound on the number of edges we removed. For (1), by the Regularity Lemma, there are at most ϵk^2 irregular pairs. Notice that for any V_i, $|V_i| \leq \frac{n}{k}$. And so each V_i has at most $\binom{\frac{n}{k}}{2}$ edges, so at most $\frac{n^2}{k^2}$ edges. Thus we remove no more than $\left(\epsilon k^2\right)\left(\frac{n^2}{k^2}\right) = \epsilon n^2$ edges.

For (2) we remove at most $k\left(\frac{n^2}{k^2}\right) \leq \frac{n^2}{k}$ edges. Notice that this number could be very large indeed if k is too small. Fortunately, the Regularity Lemma allows us to put an arbitrarily large lower bound on the value of k by choosing m.

For (3) There are $\binom{k}{2}$ total pairs and $|V_i||V_j| \leq \frac{n^2}{k^2}$. Thus we remove no more than $\binom{k}{2} a \left(\frac{n^2}{k^2}\right) < \frac{an^2}{2}$ edges.

Putting this all together we get that in total the number of edges removed is bounded above by $\left(\epsilon n^2 + \frac{n^2}{k} + \frac{1}{2}an^2\right) = \left(\epsilon + \frac{1}{k} + \frac{a}{2}\right) n^2$. Now since we chose $\epsilon + \frac{1}{m} \leq \frac{a}{2}$ it follows that the total number of edges removed is bounded above by an^2.

Let $G^{'}$ be our graph after removing these edges. Then if there is a triangle $(x, y, z) \in G^{'}$ we have that $(x, y, z) \in V_i \times V_j \times V_k$ (where the i, j, k are distinct). Furthermore we must have that (V_i, V_j), (V_i, V_k), and (V_j, V_k) are all ϵ-regular pairs with greater than $a|V_i||V_j|$ edges between them. Thus for each of the pairs $d_{ij} = \frac{e(|V_i||V_j|}{|V_i||V_j|} \geq a$.

By the counting lemma we can claim that the expected number of triangles is $(a^3)|V_i||V_j||V_k| \geq a^3 \left(\frac{n}{2k}\right)^3$ (we use the fact that $|V_i| \geq \frac{n}{2k}$ here). Therefore there are more than $\frac{a^3}{(2k)^3} n^3$ triangles present in G. Thus choose $c < \frac{a^3}{(2k)^3}$ to obtain a contradiction (we assume G has fewer than cn^3 triangles). Therfore no (x, y, z) can be in $G^{'}$, so $G^{'}$ is triangle free. $\square$

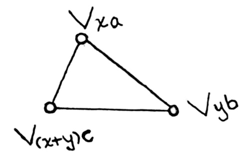
Lemma 3.2. *For every $\delta > 0$ there is an N such that $\forall\ A \subseteq N \times N$ such that $|A| \geq \delta N^2$, $\exists\ (x, y), (x + d, y), (x, y + d) \in A$ where $d \neq 0$.*

Proof. Choose N large enough. The criterion for large enough will be addressed later on in the remark that follows. Now let $A \subseteq N \times N$ such that $|A| \geq \delta N^2$. Based on A we construct a graph G as follows:

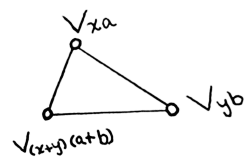
For each $a \in \{1, ..., 2N\}$ construct 3 vertices v_{xa}, v_{ya}, $v_{(x+y)a}$

The edges will be constructed as follows:
If $(a, b) \in A$ then $(v_{xa}, v_{yb}) \in E(G)$
If $(a, c - a) \in A$ then $(v_{xa}, v_{(x+y)c}) \in E(G)$
If $(c - b, b) \in A$ then $(v_{yb}, v_{(x+y)c}) \in E(G)$

Then notice the only triangles contained in G are of the form below:



Let $d = c - b - a$, for a triangle of the above form in G we have that $(a, b), (a, d + b), (d + a, b) \in A$. Thus we must show there exists a triangle in G that corresponds to some $d \neq 0$. Suppose not. Then the only triangles in G are those for which d=0, i.e.



For every $(a, b) \in A$ clearly v_{xa}, $v_{yb} \in G$ by our construction. Since $a + b \leq 2N$ we are assured that $v_{(x+y)(a+b)}$ exists. Thus for every $(a, b) \in A$ there is a triangle of the above form. Thus the number of triangles in G is exactly $|A|$. Using the fact that $\delta N^2 \leq |A| \leq N^2$ we apply the triangle removal lemma. Let $a < \frac{1}{6^2}\delta$. The choice of c is trivial and does not depend on a, so let $c = 1$. Then G contains at most $cn^3 = (6N)^3$ triangles. Thus by the triangle removal lemma it suffices to remove $a(6N)^2$ edges from G to make it triangle free.
Notice that no two triangles in A share an edge. Thus removing an edge from G removes no more than one triangle. Then one must remove exactly $|A|$ edges to make G triangle free. However, from the lemma above $a(6N)^2 < \delta N^2 < |A|$ and so we obtain a contradiction. Therefore there must be a triangle in G that corresponds to some $d \neq 0$. □

Remark 3.3. (Choice of N) In the proof above have that $|G| = 6N$ and we use the triangle removal lemma to pick $a < \frac{1}{6^2}\delta$. From the triangle removal lemma we use the Szemerédi Regularity Lemma and require that ϵ and m be chosen so that $\epsilon + \frac{1}{m} < \frac{a}{2}$ so $\epsilon + \frac{1}{m} < \frac{1}{2 \cdot 6^2}\delta$. Then from the Regularity Lemma we get an integer $N(\epsilon, m)$ so that if G has $\geq N(\epsilon, m)$ vertices the statement of the lemma is true for G and so the statement of the triangle removal lemma is true for G. So in short we need that $6N \geq N(\epsilon, m)$ where $\epsilon + \frac{1}{m} < \frac{1}{2 \cdot 6^2}\delta$.

Remark 3.4. From the lemma above and the proof it seems that it may be possible to make similar claims about the structure of A with different operations- is there anything special about addition? For instance suppose we have the following:
Potential Lemma: For every $\delta > 0$ there is an N such that $\forall\ A \subseteq N \times N$ such that $|A| \geq \delta N^2$, $\exists\ (x, y), (xd, y), (x, yd) \in A$ where $d \neq 1$.
Now suppose we go about proving this in the same manner as we proved Lemma 3.2. We construct a graph G in the same way as we did before except we replace subtraction with division. So for a given a we construct 3 vertices v_{xa}, v_{ya}, v_{xya}.

We construct the edges as follows:
If $(a,b) \in A$ then $(v_{xa}, v_{yb}) \in E(G)$
If $(a, c/a) \in A$ then $(v_{xa}, v_{(xy)c}) \in E(G)$
If $(c/b, b) \in A$ then $(v_{yb}, v_{(xy)c}) \in E(G)$

We set $d = c/ab$ so that if there is a triangle in G then $(a,b), (a, c/a), (c/b, b) \in A$ so $(a,b), (a, db), (da, b) \in A$. We want to construct this graph so that if the only triangles in G are those for which $d = 1$ then there are precisely $|A|$ triangles. To be able to make this claim we need that for every $(a,b) \in A$, $v_{(xy)ab}$ exists in G. Thus we need to construct N^2 vertices for G since the largest value ab can take on is N^2. So $n = |G| = N^2$. But now we have a problem because the triangle removal lemma gives us no contradiction- it tells us that $\forall a$ we can remove no more than aN^4 edges to make G triangle free. But $|A| < aN^4$ (a cannot depend on N) so this is trivially true.
From this exploration we can see that there is indeed something special about addition. There is a relationship between the size of the graph with the feature we are trying to find (i.e. a triangle) and the density of the corresponding feature in our set A when A is as large as possible (N^2). When we switch to multidplication, this density goes down so much that we are no longer able to use the idea.

Corollary 3.5. *For every $\delta > 0$ there is an N such that any subset $A \subseteq \{1, ..., N\}$ of size at least δN contains an arithmetic progression of length 3.*

Proof. Choose N large enough (the choice will be clear later on). Given A such that $|A| \geq \delta N$ let $B = \{(x,y) \big| x + 2y \in A\}$. Now we must determine the size of B. For each $a \in A$ there are $\lfloor \frac{a}{2} \rfloor$ xy-pairs such that $x + 2y = a$. Thus $|B| = \sum\limits_{a \in A} \lfloor \frac{a}{2} \rfloor$. In this way the size of B depends not only on the number of elements of A but the elements themselves. The smallest B could be is when A is the set $\{1, ..., \lceil \delta N \rceil\}$. In this case

$$|B| = \sum_{i=1}^{\lceil \delta N \rceil} \lfloor \tfrac{i}{2} \rfloor \geq \sum_{i=1}^{\lfloor \frac{\lceil \delta N \rceil}{2} \rfloor} 2i > 2 \sum_{i=1}^{\frac{\delta N}{3}} i = 2\left(\frac{\left(\frac{\delta N}{3} + 1\right)\frac{\delta N}{3}}{2} \right) > \frac{\delta^2 N^2}{3^2}$$

So let $\delta' = \frac{\delta}{3}$. From lemma 3.2 we can pick an N based on this δ' so that $|B| \geq \delta' N^2$ implies there exists a $d \neq 0$ such that $(x,y), (x+d, y), (x, y+d) \in B$ for some x and y. From this we get that $x + 2y, x + d + 2y$, and $x + 2y + 2d$ form an arithmetic progression in A.

□

4. Hypergraph Version of Szemerédi Regularity Lemma

We assume the reader is familiar with the definition of a 3-hypergraph. In the analogy of a graph to a hypergraph, the vertices stay the same and the edges become triples (i.e. (x,y,z) for $x, y, z \in V(G)$. newline

Definition 4.1. (Simplex) A simplex of a 3-hypergraph is a set of 4 triples: (x,y,z), (x,y,w), (x,z,w), (y,z,w). (This the analogous to a triangle in a graph).

Definition 4.2. (Quadripartite 3-Hypergraph) A quadripartite 3-hypergraph is a 3-hypergraph, H, such that one can find a partition of H's vertex set: X_1, X_2, X_3, X_4

such that there are no triples within each X_i.

Definition 4.3. (Induced Subhypergraph) Given a quadripartite 3-hypergraph, H, with vertex sets X_1, X_2, X_3, X_4, let $A_i \subset X_i$. Let $G(A_i, A_j)$ be a random bipartite graph on A_i and A_j (G is not a hypergraph!). For $1 \le i < j < k \le 4$ let G_{ijk} be the tripartite subgraph formed by the union of $G(A_i, A_j)$, $G(A_i, A_k)$, and $G(A_j, A_k)$. From this tripartite subgraph construct a 3-hypergraph H_{ijk}, where (x, y, z) is a triple in H $\iff$ x, y, z form a triangle in G. H_{ijk} is an induced subhypergraph of H.

Definition 4.4. (Expected Simplex Density) Given a quadripartite 3-hypergraph, H, let $A_i \subset X_i$. For $1 \le i < j < k \le 4$ construct H_{ijk} as above. Let $G = \cup_{ijk} H_{ijk}$. Let δ_i be such that $|A_i| = \delta_i |X_i|$. Let $\delta_{ij} = |G(A_i, A_j)|/|A_i||A_j|$. Where $|G(A_i, A_j)|$ denotes the number of edges in the random bipartite graph on A_i and A_j. Let δ_{ijk} be the relative density of H inside of H_{ijk} (i.e. the number of triples in H_{ijk} that are present in H divided by the total number of triples in H_{ijk}). The expected simplex denisty of H in G is the product of all the δ_i, δ_{ij}, and δ_{ijk}s.

Remark 4.5. The proof is omitted. However, this is where the Szemerédi Regularity Lemma plays a key role. The concept of regularity is redefined in the case of hypergraphs to a concept of "quasirandomness". The concept will not be discussed in this paper. The most important feature is that the redefined concept still allows us to make a claim about the existence of a decomposition of a hypergraph such that the components satisfy a kind of regularity. Recall the Counting Lemma- using the regularity lemma we were able to, in some sense, view the regular partitions as vertices and the density between the partitions as probability of an edge between the vertices. This allowed us to calculate the expected number of various features of the graph (the expected number of triangles was used in the proof of the triangle removal lemma). Similarly, in the hypergraph case, we define regularity in such a way that our decomposition allows us to view the density between the components as probability. Thus we are able to calculate the expected number of various features of the hypergraph- most importantly, we are able to calculate the expected simplex density.

Lemma 4.6. *(Expected Simplex Density Lemma) Let $\epsilon > 0$ and let H be a quadripartite graph with vertex sets X_1, X_2, X_3, X_4. Let $N_i = |X_i|$. It is possible to 1) remove fewer than $\epsilon|H(X_i, X_j, X_k)|$ triples of $H(X_i, X_j, X_k)$ for each i, j, k combination and 2) find a decomposition of the complete quadripartite graph $K = K(X_1, X_2, X_3, X_4)$ such that the following is satisfied: Let n denote the maximum number of partitions among all the X_is in the decomposition of K and let m be the number of bipartite graphs in any of the complete bipartite graphs between the X_i. Let H' denote the hypergraph H after the edges have been removed. Now each X_i is partitioned into at most n new vertex sets, denote them as $X_{i1}, ..., X_{in}$. Recall how we constructed H_{ijk} for each i, j, k combination in defining the expected simplex density. We do this again, except this time we construct it between each of the* partitioned *vertex sets- for instance we could construct $H_{i1,j4,k5}$ (if $n > 5$). Let G be the union of all of these induced subhypergraphs. Let σ denote the expected simplex density of H' in G. Then we have that either $\sigma = 0$ or $\sigma \ge (\epsilon/8n)^4(\epsilon/48m)^6(\epsilon/8)^4$*

and furthermore the number of simplices in H' in G differs from $\sigma N_1 N_2 N_3 N_4$ by no more than $\epsilon\sigma N_1 N_2 N_3 N_4$.

The proof is omitted.

Theorem 4.7. *For every $a > 0$ there exists $c > 0$ with the following property. Let H be any quadripartite hypergraph H with vertex sets X_1, X_2, X_3, and X_4 where the size of X_i is N_i. If H has $\leq cN_1N_2N_3N_4$ simplices then it is possible to remove fewer than $aN_iN_jN_k$ triples from each of the four induced subhypergraphs $H(X_i, X_j, X_k)$ so that H becomes simplex-free.*

Proof. The proof of this theorem follows closely to the proof given in the non-hypergraph case. Apply the Expected Simplex Density Lemma for $\epsilon = a$. Construct H' as described in the lemma. Now suppose H' contains a simplex. Let G be as described in the lemma. The expected simplex density σ is greater than 0 because each tripartite part of G contains an edge of H' (not actually sure why this part is true). Then we conclude that $\sigma \geq (\epsilon/8n)^4(\epsilon/48m)^6(\epsilon/8)^4$ where n and m are parameters that depend on ϵ. Pick $c < (\epsilon/8n)^4(\epsilon/48m)^6(\epsilon/8)^4$ to get a contradiction.

□

Theorem 4.8. *For every $\delta > 0$ there is an N such that $\forall\ A \subseteq N \times N \times N$ such that $|A| \geq \delta N^3$, $\exists (x, y, z), (x + d, y, z), (x, y + d, z), (x, y, z + d) \in A$ where $d \neq 0$*

Proof. Choose N large enough. Let $A \subseteq N \times N \times N$ such that $|A| \geq \delta N^3$. Based on A we construct a quadripartite 3-hypergraph H with vertex sets X,Y,Z, and W as follows:
For each $a \in \{1, ..., 3N\}$ construct 4 vertices $v_{xa} \in X, v_{ya} \in Y, v_{za} \in Z, v_{wa} \in W$
The triples will be constructed as follows:
If $(a, b, c) \in A$ then (v_{xa}, v_{yb}, v_{zc}) is a triple in H (note this is between $X \times Y \times Z$)
If $(a, b, e - a - b) \in A$ then (v_{xa}, v_{yb}, v_{we}) is a triple in H (between $X \times Y \times W$)
If $(e - b - c, b, c) \in A$ then $(v_{we}, v_{yb}, v_{zc}$ is a triple in H (between $W \times Y \times Z$)
If $(a, e - a - c, c) \in A$ then $(v_{xa}, v_{we}, v_{zc}$ is a triple in H (between $X \times W \times Z$)

Now suppose there is a simplex in H. Then it must be of the form $(a, b, c), (a, b, e - a - b), (e - b - c, b, c), (a, e - a - c, c)$ for some choice of $a, b, c, e \in \{1, ..., 3N\}$. This implies that $(a, b, c), (a, b, e - a - b), (e - b - c, b, c), (a, e - a - c, c) \in A$. Let $d = e - a - b - c$. Then we can rewrite the above points as $(a, b, c), (a, b, c + d), (a + d, b, c), (a, b + d, c), (a, b + d, c)$. Thus we must show there exists a simplex in H that corresponds to some $d \neq -$. Supposenot. Then the only simplices in H are those for which $d = 0$ so $e = a + b + c$. Notice that for every $(a, b, c) \in A$ we can form this simplex since $a + b + c \leq 3N$ we are assured that v_{we} exists. Thus the number of simplices in H is exactly $|A|$. Using the fact that $\delta N^3 \leq |A| \leq N^3$ we apply the simplex removal lemma. Let $a < \frac{1}{30}\delta$. The choice of c is trivial and does not depend on a, so let $c = 1$. Then since H certainly has $\leq (1)(3N)^4$ simplices it is possible to remove fewer than $a(3N)^3$ triples from each of the four induced subhypergraphs so that H becomes simplex free. But notice that no two simplices in A share a face (i.e. a triple). Thus removing a triple from H removes no more than one simplex. Then one must remove exactly $|A|$ triples to make H simplex free. However, $a(3N)^3 < \delta N^3 < |A|$ and so we obtain a contradiction. Therefore there must be simplex in H that corresponds to some $d \neq 0$.

$\square$

Corollary 4.9. *For every $\delta > 0$ there is an N such that any subset $A \subseteq \{1, ..., N\}$ of size at least δN contains an arithmetic progression of length 4.*

Proof. Choose N large enough. Given A such that $|A| \geq \delta N$ let

$$B = \left\{(x, y, z)\middle| x + 2y + 3z \in A\right\}$$

Although in this case it was more complicated a similar method can be done as was done in the previous section to show that $|B| \geq \delta' N^3$. Based on this δ' we pick on N so that by the theorem above there exists a $d \neq 0$ such that $(x, y, z), (x + d, y, z), (x, y + d, z), (x, y, z + d) \in B$. From this we get that $x + 2y + 3z, x + d + 2y + 3z, x + 2y + 2d + 3z, x + 2y + 3d + 3z$ form an arithmetic progression of length 4 in A.

$\square$

Acknowledgements. I would like to thank Professor Malliaris for her incredible mentorship over the quarter. Although not the focus of this paper, her help on understanding the meaning of the Regularity Lemma, the various applications of it, and the basic idea of multiple proofs of it made this reading course a joy. I would also like to acknowledge that the majority of the theorems in this paper come directly from the work done by Gowers.

References

[1] W.T. Gowers. Quasirandomness, counting and regularity for 3-uniform hypergraphs, *Combin. Probab. Comput.* **15** (2006), 143-184.

[2] W.T. Gowers. Hypergraph regularity and the multidimensional Szemerédi theorem, *Annals of Mathematics*, **166** (2007), 897-946.

[3] J. Komlós, M. Simonovits. Szemerédi's Regularity Lemma and Its Applications in Graph Theory, *Combinatorics* **Budapest** (1996), pp. 295-352.