

**ROTH'S THEOREM  
THE COMBINATORIAL APPROACH**

AKOS MAGYAR

We present Szemerédi's proof on the existence of 3-progressions in large sets, which appeared 15 years later than Roth's original proof, and contain elements of his general result on the existence of arbitrarily long progressions.

The key observation is to show that, if a set  $A \subset [1, N]$ ,  $\#A = \delta N$ , does not contain any 3-progression, then there is a "long" progression  $P$  on which the density of  $A$  increases at least to, say  $\delta + \delta^2/16$ . This procedure can be iterated, and if  $N$  is large enough, then after a large number  $k$  of iterations, the density of  $A$  would increase to  $\delta_k > 5/6$  on a progression  $P_k$ , which clearly implies a 3-progression.

Some preparations are helpful before we start the actual proof.

**Definition 1.** Let  $A, B$  be two finite sets of integers. The density of  $A$  in  $B$  is defined by

$$\delta(A|B) = \frac{\#(A \cap B)}{\#B}.$$

**Proposition 2.** Let  $A, B = B_1 \cup \dots \cup B_k$  be two finite sets of integers such that  $B$  is partitioned into the sets  $B_i$ . Then

$$\delta(A|B) = \delta(A|B_1) \frac{\#B_1}{\#B} + \dots + \delta(A|B_k) \frac{\#B_k}{\#B},$$

and in particular:  $\delta(A|B) \leq \max_i \delta(A|B_i)$ .

The proof of this proposition is left as an exercise, however we prove a stronger version below.

**Proposition 3.** Let  $1 \leq M < N$ ,  $\delta \leq 8/9$ , and let  $A \subset [1, N]$  such that  $\#A = \delta N$ . Assume that there exists a set  $B$ ,  $A \subset B \subset [1, N]$ , such that  $\#B \leq (1 - \delta/8)N$ . If  $B = B_1 \cup \dots \cup B_k$  is a partition, such that the number of parts  $k \leq N/M$ , then there exists a part  $B_i$  such that

$$\delta(A|B_i) \geq \delta + \delta^2/16 \quad \text{and} \quad \#B_i \geq \delta^3 M/16.$$

*Proof.* Let us call a part  $B_i$  *small* if  $\#B_i \leq \delta^3 M/16$  and *large* otherwise. Note that

$$\delta(A|B) = \frac{\#A}{\#B} \geq \frac{\delta}{1 - \delta/8} \geq \delta + \frac{\delta^2}{8}.$$

Also by Proposition 2:

$$\delta(A|B) = \sum_{\{i: B_i \text{ small}\}} \delta(A|B_i) \frac{\#B_i}{\#B} + \sum_{\{i: B_i \text{ large}\}} \delta(A|B_i) \frac{\#B_i}{\#B}.$$

However

$$\sum_{\{i: B_i \text{ small}\}} \delta(A|B_i) \frac{\#B_i}{\#B} \leq \sum_{\{i: B_i \text{ small}\}} \frac{\#B_i}{\#B} \leq \frac{N}{M} \cdot \frac{\delta^3 M/16}{\delta N} = \frac{\delta^2}{16}.$$

Hence

$$\delta + \frac{\delta^2}{16} \leq \sum_{\{i: B_i \text{ large}\}} \delta(A|B_i) \frac{\#B_i}{\#B} \leq \max_{\{i: B_i \text{ large}\}} \delta(A|B_i),$$

and this is what the proposition states. □

---

Special thanks to Liangpan Li, Shanghai Univ., for re-writing and correcting the notes.

At this point, it is enough to cover the set  $A$ , with a union of disjoint progressions  $B_1, \dots, B_k$  such that the number of them is not more than  $k \leq \frac{N}{\log \log N}$ , and moreover their union  $B$  has at most  $(1 - \delta/8)N$  elements. Indeed, then by the above proposition, one of them, say  $B_i$ , has the property that  $\#B_i \geq \delta^3 \log \log N/16$  and  $\delta(A|B_i) \geq \delta + \delta^2/16$ . Then we achieve that the density of  $A$  increases on a “long” arithmetic progression.

We have to establish a few simple facts about partitioning sets into progressions of a common difference  $d$ , whose proof is again left as an exercise.

**Definition 4.** Suppose  $d \in \mathbb{N}$  and let  $A \subset \mathbb{N}$  be a finite set. We say that  $a \in A$  and  $b \in A$  are  $d$ -equivalent in  $A$ , and write  $a \stackrel{d}{\sim} b$  if there exists a progression:

$$P = \{a, a + d, \dots, a + sd = b\} \subset A \quad \text{or} \quad P = \{b, b + d, \dots, b + sd = a\} \subset A.$$

**Proposition 5.** One has

- The relation  $\stackrel{d}{\sim}$  is an equivalence relation and the equivalence classes are maximal progressions:  $A = P_1 \cup \dots \cup P_k$  of common difference  $d$ .
- One has  $k = \#\{(A + d) \setminus A\}$ .
- The complement of  $A$  in any interval  $[n_1, n_2]$  is partitioned into at most  $k + d$  progressions.

The last key ingredient of the proof is based on the notion below.

**Definition 6.** Let  $a, d_1, \dots, d_k$  be natural numbers. A  $k$ -cube is a set of the form

$$\mathcal{M}(a : d_1, \dots, d_k) = \{a + \sum_{i=1}^k \varepsilon_i d_i : \varepsilon_i \in \{0, 1\}, \forall i = 1, \dots, k\}.$$

Note, that a 1-cube is just a pair of points  $\{a, a + d_1\}$ , and a 2-cube is of the form  $\{a, a + d_1, a + d_2, a + d_1 + d_2\}$ . In general a  $k$ -cube has  $2^k$  elements. Besides, if  $\mathcal{M}_i = \mathcal{M}(a : d_1, \dots, d_k)$ , then  $\mathcal{M}_{i+1} = \mathcal{M}_i \cup (\mathcal{M}_i + d_{i+1})$ .

**Lemma 7.** Suppose  $0 < \delta < 1$  and let  $k, N$  be natural numbers such that

$$\log \log N \geq k + \log \log(4/\delta).$$

If  $A \subset [1, N]$  and  $\#A \geq \delta N$ , then  $A$  contains a  $k$ -cube.

*Proof.* Since the number of pairs

$$\#\{(a, b) : A \ni a < b \in A\} = \frac{\#A \cdot (\#A - 1)}{2} \geq \frac{\delta^2 N(N - 1)}{4},$$

there exists a difference, say  $d_1$  for example, which is the common difference of at least  $m \geq \frac{\delta^2 N}{4}$  pairs. We list the pairs as follows:

$$d_1 = b_1 - a_1 = \dots = b_m - a_m.$$

Let  $A_1 = \{a_1, \dots, a_m\}$  and  $\delta_1 = \frac{\delta^2}{4}$ . Then

$$\#A_1 \geq \delta_1 N \quad \text{and} \quad A_1 \cup (A_1 + d_1) \subset A.$$

Similarly, there exist a sequence of common differences  $\{d_i\}_{i=2}^k$  and a sequence of sets  $\{A_i\}_{i=2}^k$  such that  $(\delta_i \doteq \frac{\delta_{i-1}^2}{4})$  for all  $i = 2, \dots, k$ :

$$\#A_i \geq \delta_i N \quad \text{and} \quad A_i \cup (A_i + d_i) \subset A_{i-1}.$$

Noting the diagram

$$\delta \rightarrow \delta_1 = \frac{\delta^2}{4} \rightarrow \delta_2 = \frac{\delta^4}{4^3} \rightarrow \delta_3 = \frac{\delta^8}{4^7} \rightarrow \dots \rightarrow \delta_k = 4 \cdot \frac{\delta^{2^k}}{4^{2^k}},$$

one immediately has  $\#A_k \geq \delta_k N \geq 4$ . Pick any element  $a^* \in A_k$ , the  $k$ -cube

$$\mathcal{M}(a^* : d_k, \dots, d_1)$$

lies in  $A$ . This proves the lemma.  $\square$

**Corollary 8.** *Suppose  $0 < \delta < 1$  and let  $N$  be natural number such that*

$$\log \log N \geq 4 + 2 \log \log(4/\delta).$$

*If  $A \subset [1, N]$  and  $\#A \geq \delta N$ , then  $A$  contains a  $k$ -cube for some  $k \geq 1 + \frac{1}{2} \log \log N$ .*

*Proof.* By the above lemma,  $A$  contains a  $k$ -cube where  $k$  is the largest integer such that

$$\log \log N \geq k + \log \log(4/\delta).$$

In another words,

$$k \geq \log \log N - \log \log(4/\delta) - 1 \geq 1 + \frac{1}{2} \log \log N.$$

□

After these preparations, we turn to Szemerédi's proof of Roth's theorem.

**Lemma 9.** *Suppose  $0 < \delta < 1$  and let  $N$  be natural number such that*

$$\log \log N \geq 4 + 2 \log \log(8/\delta).$$

*If  $A \subset [1, 4N^2]$  and  $\#A = 4\delta N^2$ , then one of the following follows:*

- *$A$  contains a 3-progression.*
- *There is a progression  $P$  such that  $\#P \geq \frac{\delta^3 \log \log N}{40}$  and  $\delta(A|P) \geq \delta + \delta^2/16$ .*

*Proof.* Assume that  $A$  does not contain any 3-progression. For  $i = 0, 1, 2, 3$ , let

$$A_i = A \cap [iN^2 + 1, (i+1)N^2] \doteq A \cap E_i.$$

If  $\#A_i \leq \delta N^2/2$  for some  $i$ , then  $\#A_j \geq (\delta + \frac{\delta}{6})N^2$  for some  $j$ . But  $E_j$  is a progression on which the density of  $A$  is increased at least to  $\delta + \frac{\delta}{6}$ , and consequently the second case is satisfied. So we assume now that  $\#A_i \geq \delta N^2/2$  for all  $i$ .

Decompose  $E_1$  into intervals of length  $N$ , one of them, say  $I$  has the property that  $\delta(A|I) \geq \delta/2$ . By Corollary 8,  $A \cap I$  contains a cube of dimension

$$k \geq 1 + \frac{1}{2} \log \log N.$$

Let us denote this cube by  $\mathcal{M}(a : d_1, \dots, d_k)$ , and for  $i = 1, \dots, k$ , let  $\mathcal{M}_i = \mathcal{M}(a : d_1, \dots, d_i)$ . Next we introduce a sequence of nondecreasing sets

$$Q_i = 2\mathcal{M}_i - A_0 \doteq \{2b - c : b \in \mathcal{M}_i, c \in A_0\} \subset [1, 4N^2].$$

Now, the sets  $\{Q_{i+1} \setminus Q_i\}_{i=1}^{k-1}$  are disjoint, hence one of them must has size

$$\#(Q_{i+1} \setminus Q_i) \leq \frac{4N^2}{k-1} \leq \frac{8N^2}{\log \log N}.$$

Let  $B = [1, 4N^2] \setminus Q_i$  and  $d = 2d_{i+1}$ . Partition  $B$  into at most  $\#(Q_{i+1} \setminus Q_i) + d$  progressions of common difference  $d$ , say for example  $B = B_1 \cup \dots \cup B_m$  as in Proposition 5.

We are now in a position to apply Proposition 3. Indeed it is evident from the construction of  $Q_i$  to see that

$$A \cap Q_i = \emptyset \quad \text{and} \quad \#Q_i \geq \#A_0,$$

from which yields

$$A \subset B \quad \text{and} \quad \#B \leq (1 - \delta/8) \cdot 4N^2.$$

It follows from Proposition 3 that there exists a part  $B_i$  for which

$$\delta(A|B_i) \geq \delta + \delta^2/16,$$

and

$$\#B_i \geq \frac{\delta^3}{16} \cdot \frac{4N^2}{m} \geq \frac{\delta^3}{16} \cdot \frac{4N^2}{\#(Q_{i+1} \setminus Q_i) + d} \geq \frac{\delta^3}{16} \cdot \frac{4N^2}{8N^2/\log \log N + 2N} \geq \frac{\delta^3 \log \log N}{40}.$$

□