THE BALOG-SZEMERÉDI THEOREM

Á. MAGYAR

We prove a result originally due to Balog and Szemerédi using the so-called *Regularity Lemma*, and later by Gowers with a much improved bound. We give a proof which uses the notion of $\varepsilon-$regularity, but not the Regularity Lemma itself, the advantage is that this gives exponential bounds, as opposed to the tower-type bounds originally obtained, however falls short of the polynomial-type bounds of Gowers. The other advantage may be that the proof is self-contained.

Theorem 1. *(Balog-Szemeredi) Let Z be an abelian group, and let $A \subseteq Z$ such that $|A| = N$ and let $0 < \delta < 1$ be fixed. Assume, that A contains many "additive quadruples", that is*

$$r_4(A) = |\{(a, b, c, d) \in A^4 : \ a + c = b + d\}| \ \geq \ \delta |A|^3$$

Then there exists a subset $B \subseteq A$ such that

$$|B| \geq c(\delta)|A| \quad \text{and} \quad |B - B| \leq c(\delta)^{-1}|B| \tag{1}$$

where: $c(\delta) = \exp\left(-C\,\delta^{-3}\,log\,(1/\delta)\right)$, $C > 0$ being some absolute constant.

Here $B - B = \{a - b : \ a \in A, \ b \in A\}$ denote the difference set of the set B. The starting idea of the proof is to assign a bipartite graph to the set A, and translate the problem into a graph theoretic settings.

The following notion will play a key role in our arguments.

Definition 1. *Let G be a bipartite graph, with vertex set $V(G) = U \cup V$ and edge set $E = E(U, V)$, and let $0 < \varepsilon \leq 1$. We say that G is $\varepsilon-$regular if the following holds.*

For every pair sets $X \subseteq U$, $Y \subseteq V$ such that $|X| \geq \varepsilon|U|$ and $|Y| \geq \varepsilon|V|$ one has that

$$|\delta(X, Y) - \delta(U, V)| \leq \varepsilon \tag{2}$$

Here $\delta(X, Y) = |E(X, Y)|/|X||Y|$ denotes the density of edges between the sets X and Y ($E(X, Y)$ being the set of edges between X and Y).

First we show that a bipartite graph contains a "large" $\varepsilon-$regular subgraph.

Proposition 1. *Let G be a bipartite graph, with vertex set $V(G) = U \cup V$ and edge set $E = E(U, V)$, and let $0 < \varepsilon \leq 1$.*

The either G is $\varepsilon-$regular, or there exists a pair of subsets: $X \subseteq U$, $Y \subseteq V$ such that $|X| \geq \varepsilon^3/3\,|U|$, $|Y| \geq \varepsilon^3/3\,|V|$ moreover

$$\delta(X,Y) \geq \delta(U,V) + \varepsilon^3/3 \tag{3}$$

Proof. If G is not $\varepsilon-$regular, then there exists $U_1 \subseteq U$, $V_1 \subseteq V$, such that $|U_1| \geq \varepsilon|U|$, $|V_1| \geq \varepsilon|V|$, and either

$$\delta(U_1,V_1) \geq \delta(U,V) + \varepsilon \qquad \text{or} \qquad \delta(U_1,V_1) \leq \delta(U,V) - \varepsilon \tag{4}$$

In the first case, taking $X = U_1$, $Y = V_1$, the proposition is proved. In the second case, let $U_2 = U \backslash U_1$, $V_2 = V \backslash V_1$, and use the fact that

$$\delta(U,V) = \sum_{i,j=1}^{2} \delta(U_i,V_j)\,\lambda_{ij} \tag{5}$$

where $\lambda_{ij} = \frac{|U_i||V_j|}{|U||V|}$. Using that $\sum_{i,j} \lambda_{ij} = 1$, one has

$$\lambda_{ii}\,(\delta(U,V) - \delta(U_1,V_1)) = \sum_{(i,j)\neq(1,1)} \lambda_{ij}\,(\delta(U_i,V_j) - \delta(U,V)) \tag{6}$$

The left side of (6) is at least ε^3, thus there exists a pair U_i, V_j such that

$$\frac{\varepsilon^3}{3} \leq \frac{|U_i|}{|U|}\,\frac{|V_i|}{|V|}\,(\delta(U_i,V_j) - \delta(U,V)) \tag{7}$$

The right side of (7) is the product of three factors each of which individually at most 1, thus by (7), each must be at least $\varepsilon^3/3$. This proves the proposition, taking $X = U_i$, $Y = V_j$.

□

Iterating the above procedure, one obtains

Lemma 1. *Let G be a bipartite graph, with vertex set $V(G) = U \cup V$ and edge set $E = E(U,V)$, and let $0 < \varepsilon \leq 1$. Then the exists an $\varepsilon-$regular pair $X \subseteq U$, $Y \subseteq V$, satisfying*

$$|X| \geq c(\varepsilon)|U|, \qquad |Y| \geq c(\varepsilon)|V|, \qquad \textit{and} \qquad \delta(X,Y) \geq \delta(U,V) \tag{8}$$

where $c(\varepsilon) = \exp\left(C\,\varepsilon^{-3} log(\varepsilon^{-1})\right)$.

Proof. Let $X_0 = U$, $Y_0 = V$. Inductively define the sets X_i, Y_i as follows. If the pair X_i, Y_i is *not* $\varepsilon-$ regular, then choose a pair of subsets X_{i+1}, Y_{i+1} by Proposition 1. By (3) one has that $\delta(X_i,Y_i) \geq \delta(U,V) + i\varepsilon^3/3$. Since the density cannot be larger than 1, the process must stop at an $\varepsilon-$ regular pair X_j, Y_j, with index $j \leq 3/\varepsilon^3$. Also by (3)

$$|X_j| \geq (\varepsilon^3/3)^j\,|U| \geq (\varepsilon^3/3)^{3/\varepsilon^3}\,|U| \geq \exp\left(C\,\varepsilon^{-3} log(\varepsilon^{-1})\right)|U|$$

and the same estimates hold for $Y_j \subseteq V$. The pair $X = X_j$, $Y = Y_j$ satisfies (8) and the Lemma is proved. $\square$

The usefulness of $\varepsilon-$ regular graphs, is that they behave like "random graphs", p.e. they contain the right number of paths of length 4 between almost any pair of points $a, b \in X$. To be more precise

Proposition 2. *Let $0 < \varepsilon \leq \delta/4$ and let X, Y be an $\varepsilon-$regular pair with edge density $\delta(X, Y) = \delta$. Then there is a set $B \subseteq X$ with $|B| \geq (1-\varepsilon)|X|$ such that for every pair of vertices $a \in B,\ b \in B$ one has*

$$N_4(a,b) := |\{(c_1, c_2, c_3) \in Y \times X \times Y :\ (a, c_1), (c_1, c_2), (c_2, c_3), (c_3, b) \in E\}| \geq \frac{\delta^4}{16}|X||Y|^2 \tag{9}$$

Note that $N_4(a, b)$ is the number of paths of length 4 connecting the points a and b, and if the graph $E(X, Y)$ were random, with edge density δ, then the expected number of such paths were $\delta^4|X||Y|^2$.

Proof. Let $B_1 = \{x \in X : n(x) \geq (\delta - \varepsilon)|Y|\}$. By definition $\delta(X \backslash B_1, Y) < (\delta - \varepsilon)$ thus (by the regularity assumption) $|B_1| \geq (1-\varepsilon)|X|$. Define the set $B_2 \subseteq Y$ the same way.

Fix $a \in B_1, b \in B_1$. Let $c_1 \in B_2$ such that $(a, c_1) \in E$, and consider the sets

$$\mathcal{N}(c_1) = \{c_2 \in X :\ (c_1, c_2) \in E\}, \quad \mathcal{N}(b) = \{c_3 \in Y :\ (c_3, b) \in E\}$$

By our construction: $|\mathcal{N}(c_1)| \geq (\delta-\varepsilon)|Y|$ and $|\mathcal{N}(b)|(\delta-\varepsilon)|X|$ thus by the $\varepsilon-$regularity of the pair X, Y the number of pairs $c_2 \in \mathcal{N}(c_1)$ and $c_3 \in \mathcal{N}(b)$ such that $(c_2, c_3) \in E$ is at least: $(\delta-\varepsilon)^3|X||Y|$.

Also, $c_1 \in \mathcal{N}(a) \cap B_2$, thus the number of possible choices for c_1 is at least: $|\mathcal{N}(a)| - |Y \backslash B_2| \geq (\delta - 2\varepsilon)|Y|$. Multiplying together the above two estimates and using that $\varepsilon \leq \delta/4$, the Proposition follows with $B = B_1$.

$\square$

Going back to Theorem 1, the idea is to assign a bipartite graph G to the set A, using the notion:

<u>Definition:</u> A pair $(a, b) \in A^2$ is called "popular", if

$$m(a-b) = |\{(c,d) \in A^2 :\ a - b = d - c\}| \ \geq \ \delta|A|^2/2 \tag{10}$$

Now let $U = V = A$ and let $E(U, V)$ be the set of popular pairs. It is easy to see that under the assumption $r_4(A) \geq \delta|A|^3$ one has

$$\delta(U, V) \geq \delta/2 \tag{11}$$

Indeed, note that

$$r_4(A) = \sum_{(a,b)\in A^2} m(a-b) = \sum_{(a,b)\ popular} m(a-b) \quad + \sum_{(a,b)\ unpopular} m(a-b) =: \sum^{1} + \sum^{2}$$

By our construction: $\sum^2 \leq \delta|A|^3/2$, thus $\sum^1 \geq \delta|A|^3/2$. Clearly $m(a-b) \leq |A|^2$ for any pair a, b, thus the number of popular differences must be at least $\delta|A|/2$.

After these preparations, it is easy to give the

Proof of Theorem 1. We use Lemma 1 with $\varepsilon = \delta/8$, to construct the $\varepsilon-$regular pair $X \subseteq A$ and $Y \subseteq A$ to the above graph G of "popular" pairs. Then we claim that the set B given in Proposition 2, will satisfy (1).

To see this, define for $c = a - b \in B - B$ the "multiplicity" function

$$m_4(c) = |\{(x_1, y_1, \ldots, x_4, y_4) \in A^8 : \ a - b = x_1 - y_1 + \ldots + x_4 - y_4\}| \tag{12}$$

Clearly

$$\sum_{c \in B-B} m_4(c) \leq |A|^8 \tag{13}$$

The crucial observation is that, by writing

$$a - b = a - c_1 + c_1 - c_2 + c_2 - c_3 + c_3 - b$$

where all 4 differences on the right side are "popular" the number of 8-tuples $(x_1, y_1, \ldots, x_4, y_4)$ such that

$$x_1 - y_1 = a - c_1,\ x_2 - y_2 = c_1 - c_2,\ x_3 - y_3 = c_2 - c_3,\ x_4 - y_4 = c_3 - b$$

is at least: $\frac{\delta^4}{16}|A|^4$. However, by Proposition 2, the number of such triples (c_1, c_2, c_3) is at least:

$\delta^4/16\,|X||Y|^2 \ \geq\ c(\delta)\,|A|^3$. Thus for any $c = a - b \in B - B$ one has

$$m_4(c) \geq c'(\delta)\,|A|^7 \tag{14}$$

Finally, note that (1) follows immediately from (13) and (14), and an easy calculation shows that the constant obtained satisfies the bound $c'(\delta) \geq \exp\left(C\,\delta^{-3} log\,(\delta^{-1})\right)$. This proves Theorem 1. $\square$