An Exposition of the Balog-Szemerédi Theorem

Ernie Croot

August 10, 2004

1 Introduction

In this paper we give the original proof of the Balog-Szemerédi Theorem, as appears in Nathanson's *Additive Number Theory: Inverse Problems and the Geometry of Sumsets*. By "original" here I mean that there is a more recent, simpler proof due to Gowers, which has the additional advantage of giving much sharper bounds on the respective constants that appear in it, as it does not appeal to the Szemerédi Regularity Lemma (which is known to give "tower-type" bounds for certain parameters in its conclusion).

Let us first state a simplified version of the theorem, which is the only version we will bother to prove in this note:

Theorem 1 *For every $\epsilon_0, \epsilon_1 > 0$, there exists $C > 0$ and $\delta_0, \delta_1 > 0$ such that the following holds: Suppose A is a finite set of integers having cardinality k, and further suppose there is a subset S of $2A$, having cardinality at least $\epsilon_0 k$, such that for every $s \in S$,*

$$\#\{(a,b) \in A \times A \ :\ a+b = s\}\ >\ \epsilon_1 k.$$

Then, if $|A| > C$, there exists a subset $A' \subseteq A$ such that

$$|A'| \geq \delta_0 k, \text{ and } |2A'| \leq \delta_1 k.$$

Note that the set S cannot be too large as

$$|S|\ \leq\ \frac{1}{\epsilon_1 k} \sum_{s \in S} \#\{a, b \in A\ :\ a+b = s\}\ \leq\ \frac{|A|^2}{\epsilon_1 k}\ =\ \frac{k}{\epsilon_1}. \qquad (1)$$

To prove Theorem 1, we will make use of the celebrated regularity lemma. First, we need to introduce some definitions. Suppose that G is an undirected graph, and that X and Y are two disjoint sets of vertices of G. Let $e(X,Y)$ denote the number of edges that connect a vertex of X to a vertex of Y. Define the density

$$d(X,Y) \ = \ \frac{e(X,Y)}{|X||Y|}.$$

Note that

$$0 \ \leq \ d(X,Y) \ \leq \ 1.$$

We now can state the regularity lemma:

Theorem 2 (Szemerédi's Regularity Lemma) *For every* $0 < \epsilon < 1$ *and* $m \geq 1$*, there exist numbers* M *and* K *such that the following holds: Suppose that* G *is a graph having vertex set* V *and edge set* E*, where* $|V| \geq K$*. Then, there is a partition of* V *into sets* $V_0, V_1, ..., V_\mu$*, where*

$$m \ \leq \ \mu \ \leq \ M,$$

which have the following remarkable properties

1. $|V_0| \leq \epsilon |V|$*. This set* V_0 *is called the exceptional set.*

2. $|V_1| = |V_2| = \cdots = |V_\mu|$*.*

3. All but at most $\epsilon\mu^2$ *of the pairs* (V_i, V_j)*,* $1 \leq i < j \leq \mu$ *are* ϵ*-regular. A pair* (V_i, V_j) *is said to be* ϵ*-regular if for every*

$$X \subseteq V_i, \quad \text{and } Y \subseteq V_j,$$

with

$$|X| \geq \epsilon |V_i|, \quad \text{and } |Y| \geq \epsilon |V_j|,$$

we have

$$|d(X,Y) - d(V_i, V_j)| \ < \ \epsilon.$$

Remark. The unfortunate thing about the regularity lemma is that the constant M can have "exponential tower" growth in terms of m and ϵ^{-1}. This is not merely an artifact of the method used to prove the theorem, but is in fact necessary, as a paper of Gowers shows.

2 Proof of the Balog-Szemerédi Theorem

2.1 Basic Strategy

The basic idea for proving the theorem will be to produce a subset $S' \subseteq S$ such that

$$|S'| \ \geq \ c_1 k,$$

and

$$|2S'| \ \leq \ c_2 k.$$

Let us now see that the mere existence of such a set S' proves the theorem: First, we claim that there exists $a \in A$ such that the number of elements $b \in A$ satisfying $a + b \in S'$ is at least $c_3 k$, where c_3 does not depend on k; this follows since there are at least $|S'|\epsilon_1 k \geq c_1 \epsilon_1 k^2$ pairs $(a, b) \in A \times A$ such that $a + b \in S'$. For such an $a \in A$, let A' be the set of all these numbers $b \in A$ such that $a + b \in S'$; thus,

$$A' + a \ \subseteq \ S', \ \ \text{and} \ \ |A'| \ \geq \ c_3 k.$$

Also,

$$|2A'| \ = \ |2A' + 2a| \ \leq \ |2S'| \ \leq \ c_2 k,$$

and the theorem is proved for $\delta_0 = c_3$ and $\delta_1 = c_2$.

2.2 Proving that $2S'$ is small

Suppose C is a set of integers satisfying

$$|C| \ \leq \ k,$$

and for which we want to prove that

$$|2C| \ \leq \ c_4 k.$$

An obvious approach to this problem is to show that each $n \in 2C$ has at least $c_5 k$ representations as $n = x_1 + x_2$, with $x_1, x_2 \in C$: For if this holds for all such n, then we can deduce that

$$|2C| \ \leq \ \frac{1}{c_5 k} \sum_{n \in 2C} r(n) \ = \ \frac{|C|^2}{c_5 k} \ \leq \ \frac{k}{c_5},$$

where $r(n)$ is the number of pairs $(x_1, x_2) \in C \times C$ such that $x_1 + x_2 = n$.

One can generalize this idea further by introducing greater flexibility into the argument, to make it easier to prove $|2C|$ is small. For example, suppose that one knows that the sumset

$$C_1 + C_2 + C_3$$

contains $2C$, where

$$|C_1|,\ |C_2|,\ \text{and}\ |C_3|\ \leq\ k.$$

Further, suppose that

$$r(n)\ \geq\ c_5 k^2,$$

where, in this context, $r(n)$ is the number of representations of $n \in 2C$ as $n = x_1 + x_2 + x_3$, where $x_i \in C_i$. Then, we will have

$$|2C|\ \leq\ \frac{1}{c_5 k^2} \sum_{n \in 2C} r(n)\ \leq\ \frac{|C_1||C_2||C_3|}{c_5 k^2}\ \leq\ \frac{k}{c_5}.$$

As good as this approach seems, without some extra knowledge about the set S', it will not prove that $|2S'|$ is small, as we would like. In the next section we will give a construction of a set S' having a lot of additional structure, and we will use this last technique for bounding $|2C|$ to bound $|2S'|$ from above.

2.3 The Connection with Graphs (no pun intended)

A common technique to prove such a combinatorial result as finding a set S' is to somehow encode the problem in terms of graphs, and in this section we will explain how to do this.

One obvious thing to try is to build a graph where the vertex set $V = A$, and where there is an edge between the vertices a_1 and a_2 if and only if $a_1 + a_2 \in S$. Denote this graph by G. To this graph G we will also give an edge coloring with $|S|$ colors, where the edge (a_1, a_2) has color $a_1 + a_2$. For a vertex set $V_0 \subseteq V$ and a color $s \in S$, by $V_0(s)$ we denote the set of all vertices $v \in V_0$ which are endpoints of an edge of color s.

Suppose that there are two vertex sets $V_1, V_2 \subseteq V$ and a color set $S' \subseteq S$, such that $|S'| \geq c_7 k$, and there is "high-connectivity" between those vertices in V_1 and in V_2 that are endpoints of edges of the colors $s \in S'$. More specifically, we suppose that:

For any $s_1, s_2 \in S'$ there are at least $c_9 k^2$ edges between $V_1(s_1)$ and $V_2(s_2)$; that is, $e(V_1(s_1), V_2(s_2)) \geq c_9 k^2$.

If we can find such a pair of vertex sets V_1 and V_2 and the color set S', then our theorem will follow rather easily from the ideas listed in Sections 2.1 and 2.2. To show that $2S'$ is small, we will show that each element $s^* \in 2S'$ has "many" representations as

$$s^* \ = \ s + a_1 + a_2, \text{ where } s \in S, \ a_1, a_2 \in A.$$

That is, we will show that there are at least $c_{10}k^2$ triples $(s, a_1, a_2) \in S \times A \times A$ such that

$$s^* \ = \ s + a_1 + a_2.$$

To see that this implies our theorem, we note that if each s^* has so many triples (s, a_1, b_2), then from the upper bound on $|S|$ in (1) we deduce

$$|2S'| \ \leq \ \frac{|S||A|^2}{c_{10}k^2} \ \leq \ \frac{k^3}{\epsilon_1 c_{10} k^2} \ = \ \frac{1}{\epsilon_1 c_{10}} k.$$

The theorem would then follow from the comments in Section 2.1.

To show that there are so many triples (s, a_1, a_2), for each $s^* \in 2S'$, fix a representation $s^* = s_1 + s_2$, where $s_1, s_2 \in S'$. Now suppose that $v_1 \in V_1(s_1)$ and $v_2 \in V_2(s_2)$ have an edge between them; as stated above, the number of such pairs $(v_1, v_2) \in V_1(s_1) \times V_2(s_2)$ with this property is at least $c_9 k^2$. By the definitions of the sets $V_1(s_1)$ and $V_2(s_2)$, associated to this pair (v_1, v_2) is a unique pair of vertices (a_1, a_2), such that there is an edge of color s_1 between v_1 and a_1, and an edge of color s_2 between v_2 and a_2; these vertices a_1 and a_2 may not lie in V_1 or V_2, which does not matter for our argument. Thus, associated to s^* is a set of at least $c_9 k^2$ distinct triples

$$(s, a_1, a_2) \ = \ (v_1 + v_2, a_1, a_2) \ \in \ S \times A \times A,$$

such that

$$s + a_1 + a_2 \ = \ (v_1 + a_1) + (v_2 + a_2) \ = \ s_1 + s_2 \ = \ s^*,$$

which is just what we wanted to show.

2.4 Regularity Lemma and Proof of the Existence of V_1, V_2, and S'

In this section we will finish the proof of the theorem, by showing how the regularity lemma can be used to construct the vertex sets V_1 and V_2 and the color set S'.

Let $\epsilon > 0$ be as small as needed, and $m \geq 1$ be as large as needed, in order for our argument to work (that is, $\epsilon > 0$ and $m \geq 1$ will be chosen later). Then, we invoke the Szemerédi regularity lemma, and apply it to our graph G, while supposing that k is large enough for the lemma to apply for our particular choice of ϵ and m. Thus, we have a partition of the vertex set of G into $V_0, V_1, ..., V_\mu$ satisfying the conclusion of the regularity lemma.

Now we remove edges from the graph G, to produce a graph G', in four different phases:

1. Delete all the edges connected to vertices in the exceptional set V_0. As V_0 has at most ϵk vertices, this step removes at most ϵk^2 edges from G.

2. If a pair of vertex sets (V_i, V_j) is not ϵ-regular, then we delete all the edges between them. As there are at most $\epsilon \mu^2$ pairs of vertex sets (V_i, V_j) which are not regular, and since $|V_i| = |V_j| \leq k/\mu$ for $1 \leq i, j \leq \mu$, this step removes at most the following number of edges

$$\sum_{\substack{1 \leq i,j \leq \mu \\ (V_i,V_j) \text{ not regular}}} e(V_i, V_j) \ < \sum_{\substack{1 \leq i,j \leq \mu \\ (V_i,V_j) \text{ not regular}}} (k/\mu)^2 \ < \ \epsilon k^2.$$

3. For any vertex set V_i and any color s such that $|V_i(s)| < \epsilon |V_i|$ we delete all edges of color s with an endpoint in V_i. The total number of edges deleted, for all the vertex sets V_i, and for all colors s, is at most

$$\sum_{i=1}^{\mu} \sum_{s \in S} \epsilon |V_i| \ \leq \ \epsilon k |S| \ \leq \ \frac{\epsilon k^2}{\epsilon_1}.$$

This last inequality follows from our upper bound on $|S|$ provided by (1).

4. Finally, we delete all the edges from V_i to V_i, for each $i = 1, ..., \mu$. The number of edges deleted here is at most

$$\sum_{i=1}^{\mu} |V_i|^2 \ \leq \ \mu \left(\frac{k}{\mu} \right)^2 \ = \ \frac{k^2}{\mu}.$$

The total number of edges in G is at least

$$\sum_{s\in S} \#\{\text{edges in } G \text{ of color } s\} \ \geq\ |S|\epsilon_1 k \ \geq\ \epsilon_0\epsilon_1 k^2.$$

And, the total number of edges we removed is at most

$$\left(2\epsilon + \frac{\epsilon}{\epsilon_1} + \frac{1}{\mu}\right) k^2,$$

which can be made to be as small a proportion of the $\geq \epsilon_0\epsilon_1 k^2$ edges in G as we like, by taking $\epsilon > 0$ sufficiently small, and $m \geq 1$ sufficiently large. In particular, we can choose $\epsilon > 0$ and $m \geq 2$ so that

$$\#\{\text{edges in } G'\} \ \geq\ \frac{\epsilon_0\epsilon_1}{2} k^2.$$

Since there are at most

$$\binom{\mu}{2} < \frac{1}{2}\mu^2$$

pairs of vertex sets (V_i, V_j), there must be a pair with at least $c_{10}k^2$ edges (in G') between them, where $c_{10} = \epsilon_0\epsilon_1/\mu^2$. Such a pair (V_i, V_j) is necessarily ϵ-regular in the original graph G, for edges between non-regular vertex sets get deleted by step 2 above.

Relabeling the vertex sets as needed, we can assume that $i = 1$ and $j = 2$ so that the pair under consideration is (V_1, V_2). We now let S' be the set of all colors of edges (in G') between V_1 and V_2. To bound $|S'|$ from below, pick a vertex $v \in V_1$ having at least the average number of edges into V_2. As V_1 has at most k vertices, we conclude that there are at least $c_{10}k$ edges form v into V_2, and each of these edges must be of a different color, which gives that

$$|S'| \ \geq\ c_{10}k.$$

Now, from the way that we have deleted edges from G to produce G', and in particular, step 3 of the deletion process, we have that $|V_1(s)| \geq \epsilon|V_1|$ for any color $s \in S'$; bear in mind that not all these edges will have their other endpoint in V_2. We similarly have that $|V_2(s)| \geq \epsilon|V_2|$. Let $s_1, s_2 \in S'$. From the fact that the pair (V_1, V_2) is ϵ-regular (in the original graph G), we have that

$$d(V_1(s_1), V_2(s_2)) \ > \ d(V_1, V_2) - \epsilon.$$

(Here and below the quantities d and e refer to the original graph G, not to G'.) Now,

$$d(V_1, V_2) \ = \ \frac{e(V_1, V_2)}{|V_1||V_2|} \ \geq \ \frac{\epsilon_0 \epsilon_1 k^2}{\mu^2 |V_1||V_2|} \geq \epsilon_0 \epsilon_1.$$

By taking $\epsilon > 0$ sufficiently small, we will have that

$$d(V_1(s_1), V_2(s_2)) \ > \ \epsilon_0 \epsilon_1 - \epsilon \ > \ \epsilon.$$

It follows that

$$e(V_1(s_1), V_2(s_2)) \ > \ \epsilon |V_1(s_1)||V_2(s_2)| \ > \ \epsilon(\epsilon|V_1|)(\epsilon|V_2|) \ > \ \epsilon^3 |V_1||V_2|.$$

To bound this from below, we note for $\epsilon < 1/2$ that there are at least $k/2$ vertices in $V_1 \cup \cdots \cup V_\mu$ (because $|V_0| < \epsilon k$, and $|V_1| = |V_2| = \cdots = |V_\mu|$), which implies

$$|V_1|, \ |V_2| \ \geq \ \frac{k}{2\mu};$$

and so,

$$e(V_1(s_1), V_2(s_2)) \ > \ \frac{\epsilon^3 k^2}{4\mu^2}.$$

Thus, we have proved that the vertex sets V_1 and V_2 have the requisite properties with

$$c_9 \ = \ \frac{\epsilon^3}{4\mu^2},$$

for $\epsilon > 0$ sufficiently small; and so, the theorem follows.