THE (GOWERS-)BALOG-SZEMERÉDI THEOREM: AN EXPOSITION

VSEVOLOD F. LEV

This is an expository note not intended for publication

Abstract. We briefly discuss here the Balog-Szemerédi theorem, compare its different versions and prove, following Gowers, the strongest one — also due to Gowers.

1. Discussion

For a finite set A of elements of an abelian group and a group element s, by $\nu_A(s)$ we denote the number of representations of s as a sum of two elements of A:

$$\nu_A(s) = \#\{(a', a'') \in A \times A \colon s = a' + a''\}.$$

We write $2A = \{a' + a'' \colon a', a'' \in A\}$, the set of all elements s with $\nu_A(s) > 0$.

Our motivation will be clear from the following simple lemma.

Lemma 1. *Suppose that $|2A| \le C|A|$ with some real $C > 0$. Then*

(i) *there are at least $|A|/2$ elements $s \in 2A$ with $\nu_A(s) \ge |A|/(2C)$;*

(ii) *if W is the set of all pairs $(a', a'') \in A \times A$ with $\nu_A(a'+a'') \ge |A|/(2C)$, then $|W| \ge |A|^2/(4C)$ and furthermore there are at most $2C|A|$ distinct sums of the form $a' + a''$ with $(a', a'') \in W$;*

(iii) *the number of solutions of the equation $a_1 + a_2 = a_3 + a_4$ in the variables $a_1, a_2, a_3, a_4 \in A$ is at least $|A|^3/C$.*

Proof. Write $S := \{s \in 2A \colon \nu_A(s) \ge |A|/(2C)\}$. Then

$$|A|^2 = \sum_{s\in 2A} \nu_A(s) = \sum_{s\in S} \nu_A(s) + \sum_{s\in 2A\setminus S} \nu_A(s) \le |S||A| + |2A||A|/(2C)$$

(the sum over $s \in S$ contains $|S|$ summands not exceeding $|A|$, the sum over $s \in 2A \setminus S$ contains at most $|2A|$ summands not exceeding $|A|/(2C)$). By the assumption we have $|2A||A|/(2C) \le |A|^2/2$ implying $|S| \ge |A|/2$, as claimed in (i).

To prove (ii) we notice that

$$|W| = \sum_{s\in S} \nu_A(s) \ge |S||A|/(2C) \ge |A|^2/(4C)$$

and that the number of distinct sums of the form $a'+a''$ with $(a',a'')\in W$ is

$$|S| \le \frac{1}{|A|/(2C)} \sum_{s\in S} \nu_A(s) \le \frac{1}{|A|/(2C)} \sum_{s\in 2A} \nu_A(s) = 2C|A|.$$

Finally, (iii) is established once we observe that the number of solutions of the equation in question is

$$\sum_{s\in 2A} (\nu_A(s))^2 \ge \frac{1}{|2A|}\Big(\sum_{s\in 2A}\nu_A(s)\Big)^2 = \frac{|A|^4}{|2A|} \ge |A|^3/C.$$

$\square$

The Balog-Szemerédi theorem is an assertion "inverse" to the lemma above. In fact, it is not true that if A satisfies conditions of the sort (i)–(iii) then A has the small doubling property: consider, for instance, sets consisting of an arithmetic progression and a number of "sporadic" elements. Balog and Szemrédi have shown, however, that if (i)–(iii) hold then there is a large *subset* $A_0 \subseteq A$ with small doubling.

In the light of the discussion above it is not surprising that the assumptions of the Balog-Szemerédi theorem can be stated in several equivalent forms. More precisely, consider the following three conditions (depending on positive real parameters $\varepsilon,\delta,\tau,K$, and γ).

- **C1**(ε,δ)**:** there is a set $S\subseteq 2A$ such that $|S|\ge\varepsilon|A|$ and $\nu_A(s)\ge\delta|A|$ for any $s\in S$;
- **C2**(τ,K)**:** there is a set $W\subseteq A\times A$ such that $|W|\ge\tau|A|^2$ and the number of distinct sums $a'+a''$ with $(a',a'')\in W$ is at most $K|A|$;
- **C3**(γ)**:** the number $T(A)$ of solutions of the equation $a_1+a_2=a_3+a_4$ in the variables $a_1,a_2,a_3,a_4\in A$ satisfies $T(A)\ge\gamma|A|^3$.

These conditions are essentially equivalent: we leave it to the reader to verify that C1(ε,δ) implies C2$(\varepsilon\delta,1/\delta)$; furthermore, C2$(\tau,K)$ implies C3(τ^2/K); and finally, C3(γ) implies C1$(\gamma/2,\gamma/2)$. Once equivalence is established we can switch freely between the conditions. In practice we prefer to use C3(γ) which depends on just one parameter, and from now on we adopt $T(A)$ as a standard notation.

It is possible to re-state conditions C1 and C2 so that they will depend on just one parameter, too. It is also possible to consider sumsets of the form $A+B := \{a+b\colon a\in A,\ b\in B\}$ instead of $2A$, and indeed the most general form of the Balog-Szemerédi theorem addresses this situation. We will not discuss this below, however.

The "original" (pre-Gowers) version of the Balog-Szemerédi theorem is as follows.

Theorem 1. *Let A be a finite non-empty set of elements of an abelian group and suppose that $T(A) \geq \gamma|A|^3$, where $\gamma > 0$ is a real number. Then there exists a subset $A_0 \subseteq A$ satisfying $|A_0| \geq c|A|$ and $|2A_0| \leq C|A_0|$ with positive constants c and C depending only on γ.*

For a nicely presented classical proof of Theorem 1 (with the assumptions in the form C1) we refer the reader to [C04]. The problem with the classical proof is that it is based on the Szemerédi regularity lemma, hence the dependence of c and C on γ arising from this proof is very poor; more precisely tower-like, which in practice means ineffective. In contrast, Gowers was able to find a proof which does not use the regularity lemma and leads to "good" relation between c and C, on the one hand, and γ on the other hand. His result is also somewhat stronger than the original Balog-Szemerédi theorem.

Theorem 2 (Gowers, [G98])**.** *Let A be a finite non-empty set of elements of an abelian group and suppose that $T(A) \geq \gamma|A|^3$, where $\gamma > 0$ is a real number. Then there exists a subset $A_0 \subseteq A$ satisfying $|A_0| \geq (\gamma^2/40)|A|$ such that for any $a_1, a_2 \in A_0$ the number of solutions of the equation*

$$a_1 - a_2 = x_1 + x_2 + x_3 + x_4 - y_1 - y_2 - y_3 - y_4$$

in the variables $x_i, y_i \in A$ $(i = 1, \dots, 4)$ is at least $2^{-28}\gamma^{10}|A|^7$. Consequently, we have $|A_0 - A_0| < 2^{28}\gamma^{-10}|A|$.

We note that the "consequently part" of the theorem (to which we will never return again) is almost immediate. There are totally $|A|^8$ expressions of the form $x_1 + \dots - y_4$, and by the first assertion any element of the difference set $A_0 - A_0$ "eats up" at least $2^{-28}\gamma^{10}|A|^7$ expressions; thus the number of elements of the difference set is at most $|A|^8/(2^{-28}\gamma^{10}|A|^7)$.

It is worth pointing out that the conclusion of Theorem 2 deals with the difference set $A_0 - A_0$ rather than the sumset $2A_0$. However, the cardinalities of the two sets are known to be tightly related; in particular, by a well-known lemma of Ruzsa from $|A_0 - A_0| < C|A_0|$ it follows that $|2A_0| < C^2|A_0|$.

2. Proof of the (Gowers-)Balog-Szemerédi theorem

Suppose that we are given subsets $A_1, \dots, A_n \subseteq A$ of cardinality at least $\delta|A|$ each, with some $\delta > 0$. It is easily seen then that the average intersection $A_i \cap A_j$ has at least $\delta^2|A|$ elements. The following lemma shows that, indeed, one can select a "large" group of subsets so that "almost all" pairs of selected subsets have intersection of about expected size.

Lemma 2. *Let A be a finite non-empty set of cardinality $m = |A|$ and suppose that $A_1, \dots, A_n \subseteq A$ satisfy $|A_i| \ge \delta m$ $(i = 1, \dots, n)$, where $\delta > 0$ is a real number. Then there is a set of indices $I \subseteq [1, n]$ such that $|I| \ge \delta n/2$ and*

$$\#\{(i,j) \in I \times I \colon |A_i \cap A_j| \le 0.03\delta^2 m\} < \frac{1}{25}\,|I|^2.$$

Remark. There is the usual trade-off between the constants 0.03, 1/25, and $\delta/2$ (in $|I| \ge \delta n/2$); for instance, 0.03 can be replaced by any value, smaller than one. One can obtain slightly better constants modifying the argument as follows: instead of a random element $a \in A$ consider a random k-tuple $a = (a_1, \dots, a_k) \in A^k$ and define $I_a = \{i \in [1, n] \colon a_1, \dots, a_k \in A_i\}$. Indeed, this is the how the argument runs in Gowers' original proof.

Proof of Lemma 2. For $a \in A$ let $I_a = \{i \in [1, n] \colon a \in A_i\}$; thus $i \in I_a$ if and only if $a \in A_i$, and

$$\sum_{i,j=1}^{n} |A_i \cap A_j| = \sum_{a \in A} |I_a|^2 \ge \frac{1}{m}\Big(\sum_{a \in A} |I_a|\Big)^2 = \frac{1}{m}\Big(\sum_{i=1}^{n} |A_i|\Big)^2 \ge \delta^2 m n^2$$

(showing that the average intersection $A_i \cap A_j$ has at least $\delta^2 m$ elements).

We prove that one can take $I = I_a$ for some a. For this, choose $a \in A$ at random. Then

$$\begin{aligned}\mathsf{E}|I_a|^2 = \mathsf{E}\,\#\{(i,j) \colon i, j \in I_a\} &= \mathsf{E}\#\{(i,j) \colon a \in A_i \cap A_j\} \\ &= \sum_{i,j=1}^{n} \mathsf{P}\{a \in A_i \cap A_j\} = \sum_{i,j=1}^{n} m^{-1}|A_i \cap A_j| \ge \delta^2 n^2;\end{aligned}$$

on the other hand,

$$\begin{aligned}\mathsf{E}\#\{(i,j) \in I_a \times I_a \colon |A_i \cap A_j| \le 0.03\delta^2 m\} &= \sum_{\substack{i,j=1 \\ |A_i \cap A_j| \le 0.03\delta^2 m}}^{n} \mathsf{P}\{i, j \in I_a\} \\ = \sum_{\substack{i,j=1 \\ m^{-1}|A_i \cap A_j| \le 0.03\delta^2}}^{n} \mathsf{P}\{a \in A_i \cap A_j\} &= \sum_{\substack{i,j=1 \\ m^{-1}|A_i \cap A_j| \le 0.03\delta^2}}^{n} m^{-1}|A_i \cap A_j| \le 0.03\delta^2 n^2.\end{aligned}$$

Therefore,

$$\mathsf{E}\Big\{|I_a|^2 - 25\,\#\{(i,j) \in I_a \times I_a \colon |A_i \cap A_j| \le 0.03\delta^2 m\}\Big\} \ge \tfrac{1}{4}\,\delta^2 n^2,$$

and there exists $a \in A$ such that

$$|I_a| \geq \frac{1}{2}\,\delta n,$$
$$\#\{(i,j) \in I_a \times I_a \colon |A_i \cap A_j| \leq 0.03\delta^2 m\} < \frac{1}{25}\,|I_a|^2.$$

□

To make Lemma 2 easier to apply we restate it in a slightly different form.

Lemma 2′. *Let A and B be two finite non-empty sets. Write $m := |A|$ and suppose that to any $b \in B$ there corresponds a subset $N(b) \subseteq A$ of cardinality $|N(b)| \geq \delta m$, where δ is a positive real number. Then there exists $B' \subseteq B$ with $|B'| \geq \delta|B|/2$ such that*

$$\#\{(b',b'') \in B' \times B' \colon |N(b') \cap N(b'')| \leq 0.03\delta^2 m\} < \frac{1}{25}\,|B'|^2.$$

We need a simple graph-theoretic lemma.

Lemma 3. *Let G be a graph (possibly, with loops) on the vertex set V of average degree $\bar{d} \geq (1-\lambda)|V|$. Then V contains at least $(1-\sqrt{\lambda})|V|$ vertices of degree greater than $(1-\sqrt{\lambda})|V|$:*

$$\#\{v \in V \colon \deg(v) > (1-\sqrt{\lambda})|V|\} \geq (1-\sqrt{\lambda})|V|.$$

Proof. If k is the number of vertices $v \in V$ of degree $\deg(v) > (1-\sqrt{\lambda})|V|$, then

$$\begin{aligned}\bar{d}|V| &= \sum_{v\colon \deg(v)\leq(1-\sqrt{\lambda})|V|} \deg(v) \quad + \sum_{v\colon \deg(v)>(1-\sqrt{\lambda})|V|} \deg(v) \\ &\leq (1-\sqrt{\lambda})|V|(|V|-k) + |V|k,\end{aligned}$$

whence

$$(1-\lambda)|V| \leq (1-\sqrt{\lambda})(|V|-k) + k = (1-\sqrt{\lambda})|V| + \sqrt{\lambda}k$$

and therefore

$$k \geq (1-\sqrt{\lambda})|V|.$$

□

Combining Lemmas 2′ and 3 we get

Lemma 4. *Let A and B be as in Lemma 2′. Then there exist subsets $B_0 \subseteq B' \subseteq B$ with $|B_0| \geq 4|B'|/5 > 2\delta|B|/5$ and such that for any $b_0 \in B_0$ we have*

$$\#\{b' \in B' \colon |N(b_0) \cap N(b')| \geq 0.03\delta^2 m\} > \frac{4}{5}\,|B'|.$$

Proof. Find B' as in Lemma $2'$ and construct a graph on the vertex set $\{N(b')\}_{b'\in B'}$, joining $N(b')$ and $N(b'')$ if $|N(b')\cap N(b'')|\ge 0.03\delta^2 m$. (The case $b'=b''$ is not excluded so that each $N(b')$ is joined to itself.) The number of non-loop edges of this graph is more than $\binom{|B'|}{2}-\frac{|B'|^2}{50}$, the average degree is more than $24|B'|/25$, and applying Lemma 3 with $\lambda=1/25$ we conclude that there is a set B_0 of at least $4|B'|/5$ elements such that if $b_0\in B_0$, then $N(b_0)$ is adjacent to more than $4|B'|/5$ sets $N(b')$ with $b'\in B'$. $\square$

Proof of the Gowers-Balog-Szemerédi theorem. Write $m:=|A|$. Let $\nu_A^-(d)$ denote the number of representations of the group element d as $d=a_1-a_2$ with $a_1,a_2\in A$. We say that d is a popular difference if $\nu_A^-(d)\ge\gamma m/2$, and we let $D:=\{d\in A-A\colon \nu_A^-(d)\ge\gamma m/2\}$, the set of all popular differences.

Consider the graph G on the vertex set A, in which a_1 and a_2 are adjacent if and only if $a_1-a_2\in D$ (the case $a_1=a_2$ not excluded). The average degree of G is

$$\begin{aligned}\bar d &= \frac{1}{m}\sum_{a\in A}\deg(a)\\ &= \frac{1}{m}\sum_{a\in A}\#\{a'\in A\colon a'-a\in D\}\\ &= \frac{1}{m}\#\{(a,a')\in A\times A\colon a'-a\in D\}\\ &= \frac{1}{m}\sum_{d\in D}\nu_A^-(d),\end{aligned}$$

and to estimate the sum at the right we observe that

$$\begin{aligned}\gamma m^3\le T(A)=\sum_{d\in A-A}(\nu_A^-(d))^2 &= \sum_{d\in(A-A)\setminus D}(\nu_A^-(d))^2+\sum_{d\in D}(\nu_A^-(d))^2\\ &\le \frac12\gamma m\sum_{d\in A-A}\nu_A^-(d)+m\sum_{d\in D}\nu_A^-(d)=\frac12\gamma m^3+m\sum_{d\in D}\nu_A^-(d).\end{aligned}$$

Therefore $\sum_{d\in D}\nu_A^-(d)\ge\gamma m^2/2$ whence $\bar d\ge\gamma m/2$ and by Lemma 3 as applied with $\lambda=1-\gamma/2$, there is a subset $B\subseteq A$ such that $|B|\ge(1-\sqrt{1-\gamma/2})m>\gamma m/4$ and $\deg(b)>\gamma m/4$ for any $b\in B$. Thus if $N(b)$ denotes the neighborhood of b in G (including b itself), then $|N(b)|>\gamma m/4$ $(b\in B)$.

We now apply Lemma 4 to the system of sets $N(b)\subseteq A$ $(b\in B)$ with $\delta=\gamma/4$ to find two subsets $A_0\subseteq A'\subseteq B$ such that

(i) $|A'|\ge(\gamma/4)|B|/2\ge\gamma^2 m/32$, and $|A_0|\ge 4|A'|/5\ge\gamma^2 m/40$;

(ii) for any $a_0 \in A_0$ we have $\#\{a' \in A' \colon |N(a_0) \cap N(a')| \geq 0.03(\gamma/4)^2 m\} > \frac{4}{5}\,|A'|$.

We claim that A_0 possesses the property in question. Indeed, fix $a_1, a_2 \in A_0$ and notice that

$$\#\{a' \in A' \colon |N(a_i) \cap N(a')| \geq 0.03(\gamma/4)^2 m \text{ for } i = 1, 2\} > \frac{3}{5}\,|A'|.$$

Choose one of these values of a'. For any $a \in N(a_1) \cap N(a')$ both $a_1 - a$ and $a' - a$ are popular differences, yielding at least $(\gamma m/2)^2$ representations $a_1 - a' = (x_1 - y_1) - (x_1' - y_1')$ with

$$x_1, y_1, x_1', y_1' \in A,\ x_1 - y_1 = a_1 - a,\ x_1' - y_1' = a' - a.$$

The total number of possible values of a is $|N(a_1) \cap N(a')| \geq 0.03(\gamma/4)^2 m$, leading to at least

$$0.03(\gamma/4)^2 m\,(\gamma m/2)^2 = 0.03 \cdot 2^{-6}\gamma^4 m^3$$

representations $a_1 - a' = (x_1 - y_1) - (x_1' - y_1')$. (Notice that a' and a can be recovered from any such representation, hence all these representations are pairwise distinct.) Similarly, there are at least $0.03 \cdot 2^{-6}\gamma^4 m^3$ representations $a_2 - a' = (x_2 - y_2) - (x_2' - y_2')$ with $x_2, y_2, x_2', y_2' \in A$. Combining, we get at least $0.03^2 \cdot 2^{-12}\gamma^8 m^6$ distinct representations

$$a_1 - a_2 = (x_1 - y_1) - (x_1' - y_1') - (x_2 - y_2) - (x_2' - y_2')$$

and each of them determines a' uniquely. Using all possible a' we get at least

$$0.03^2 \cdot 2^{-12}\gamma^8 m^6 \cdot \frac{3}{5}\,|A'| > 2^{-28}\gamma^{10} m^7$$

representations, as wanted. □

References

[BS94] A. Balog, E. Szemerédi, A statistical theorem of set addition, *Combinatorica* **14** (1994), 263–268.

[C04] E. Croot, An exposition of the Balog-Szemerédi theorem, *an expository note available at http://www.math.gatech.edu/ ecroot/balog.pdf.*

[G98] W.T. Gowers, A new proof of Szemerédi's theorem for arithmetic progressions of length four, *Geometric And Functional Analysis* **8** (1998).

Department of Mathematics, University of Haifa at Oranim, Tivon 36006, Israel.
E-mail address: seva@math.haifa.ac.il