THEORY OF COMPUTING LIBRARY
GRADUATE SURVEYS 3 (2011), pp. 1–15
www.theoryofcomputing.org

Selected Results in Additive Combinatorics: An Exposition

Emanuele Viola*

Received: February 3, 2009; revised: May 6, 2011; published: May 15, 2011.

Abstract: We give a stripped-down, self-contained exposition of selected results in additive combinatorics over the vector space $\mathbb{F}_2^n$, leading to the result by Samorodnitsky (STOC 2007) stating that linear transformations are efficiently testable. In particular, we prove the theorems known as the Balog-Szemerédi-Gowers theorem (Combinatorica 1994 and GAFA 1998) and the Freiman-Ruzsa theorem (AMS 1973 and Astérisque 1999).

ACM Classification: 05D99

AMS Classification: F.1.2, F.2.2

Key words and phrases: Additive combinatorics, linearity testing

1 Introduction

Additive combinatorics is a fascinating area of mathematics that has found several applications in theoretical computer science. In addition to the book by Tao and Vu [20], a number of expositions of various results in additive combinatorics are now available. See, for example, the survey by Trevisan [21] and the pointers therein.

In this survey we aim to provide a self-contained, friendly introduction to additive combinatorics. We cover a few selected results, stripped down to the minimum needed to obtain the following result by Samorodnitsky [17] that linear transformations are efficiently testable.

*Supported by NSF grant CCF-0845003. This work was partially done while the author was at the School of Mathematics, Institute for Advanced Study, Princeton, NJ, 08540, supported by NSF grant CCR-0324906. A preliminary version appeared in [22]. Email: `viola@ccs.neu.edu`

 DOI: 10.4086/toc.gs.2011.003

Theorem 1.1 (Testing linear transformations (Samorodnitsky))**.** *For all $\varepsilon > 0$ there is $\varepsilon' > 0$ such that for all sufficiently large n and all functions $f : \mathbb{F}_2^n \to \mathbb{F}_2^n$ the following holds. If*

$$\Pr_{x,x' \in \mathbb{F}_2^n}[f(x) + f(x') = f(x+x')] \geq \varepsilon\,,$$

then there is an $n \times n$ matrix M such that

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) = Mx] \geq \varepsilon'\,.$$

In the statement of this theorem, and throughout the remainder of the paper, $\mathbb{F}_2$ denotes the set $\{0,1\}$ with mod 2 arithmetic.

Theorem 1.1 shows that the test "pick $x, x' \in \mathbb{F}_2^n$ uniformly at random, and accept if $f(x) + f(x') = f(x+x')$" is useful to check if a function $f : \mathbb{F}_2^n \to \mathbb{F}_2^n$, given as a black-box, is close to a linear transformation. Indeed, if f is a linear transformation, i. e., $f(x) = Mx$ for an $n \times n$ matrix M, then clearly the test always accepts; while on the other hand if the test accepts with probability ε then the theorem guarantees that there is a linear transformation (given by M) that agrees with f on an ε' fraction of the inputs.

This test was first proposed by Blum, Luby, and Rubinfeld [5], who analyze it for functions between finite groups. Bellare et al. [4] later refine the analysis in the case $f : \mathbb{F}_2^n \to \mathbb{F}_2$. The proofs in [5, 4] appear to break down in the setting of Theorem 1.1 where the range of f is $\mathbb{F}_2^n$.

The proof of Theorem 1.1 gives an exponential dependence of ε' on ε. The "polynomial Freiman-Ruzsa conjecture" states that ε' is polynomial in ε [10].

To introduce the subject of additive combinatorics and motivate the following sections, let us now informally see how the property-testing result in Theorem 1.1 follows from some known results in the subject, which will be presented along the way.

Proof idea for Theorem 1.1. We are interested in the additive combinatorics of the *graph* A of the function f:

$$A := \{(x, f(x)) : x \in \mathbb{F}_2^n\} \subseteq \mathbb{F}_2^{2n}\,.$$

The approach to prove the theorem is to show that A is *approximately* a linear space. This approach is motivated by the observation that if A were *exactly* a linear space, then f would be a linear transformation, because in this case

$$(x, f(x)) + (x', f(x')) = (x+x', f(x) + f(x')) \in A = \{(x, f(x)) : x \in \mathbb{F}_2^n\}\,,$$

and so $f(x) + f(x')$ must equal $f(x+x')$.

We start by noting that our assumption can be written as

$$\Pr_{a,a' \in A}[a + a' \in A] \geq \varepsilon\,. \tag{1.1}$$

Next, we apply our first result in additive combinatorics, namely the *Balog-Szemerédi-Gowers (BSG) theorem* [2, 9]. This theorem states that if a set A satisfies (1.1) then it contains a large subset that is nearly closed under addition. More formally, defining $2S := \{a+a' : a,a' \in S\}$, the BSG theorem says that there is a set $A' \subseteq A$ of large size $|A'| \approx |A|$ such that

$$|2A'| \approx |A'|. \tag{1.2}$$

(From Equation (1.1) we cannot in general conclude that $|2A| \approx |A|$, which motivates considering the subset $A' \subseteq A$.)

At this point we apply our second result in additive combinatorics, namely *Ruzsa's theorem* [16], which is a finite-field analogue of an older theorem due to Freiman [8]. This theorem says that if a set A' satisfies (1.2) then it is approximately a linear space. Specifically, denoting by $\operatorname{span}(A')$ the vector space spanned by elements of A', Ruzsa's theorem states that

$$|\operatorname{span}(A')| \approx |A'|. \tag{1.3}$$

In other words, Ruzsa's theorem says that if linear combinations of length 2 (i. e., $2A'$) do not buy much size, then neither do linear combinations of arbitrary length (i. e., $\operatorname{span}(A')$).

Finally, even though A' may not be a linear space, from (1.3) one can still draw the conclusion that f is close to a linear transformation, thus concluding the proof of the theorem. □

Before discussing how this exposition is organized, we stress that it focuses on additive combinatorics in the additive group of the vector space $\mathbb{F}_2^n$. This choice is motivated by the importance of this space to computer science, and by the fact that the proofs of the relevant results in additive combinatorics appear to be cleanest over $\mathbb{F}_2^n$. More work is needed to extend these results to more general domains mainly because of the need to take care of the signs (as discussed in [20], for example). With that more work, one can extend these results as follows. An analogue of Theorem 1.1 holds over $\mathbb{F}_p^n$ for any fixed prime p, see [17, Theorem 4.1]. The BSG theorem holds over any abelian group, see [20, Theorem 2.29]. The Freiman-Ruzsa theorem holds over any abelian group with bounded torsion, see [20, Theorem 5.27].

We also mention that Green and Tao [12] have recently given a new direct proof of the combination of the BSG and Ruzsa theorems over $\mathbb{F}_2^n$ using Fourier analysis. Going back to the proof of Theorem 1.1, their result goes directly from (1.1) to (1.3).

Organization. After some preliminaries in Section 2, we prove the BSG theorem in Section 3. In Section 4 we prove Ruzsa's theorem. In Section 5 we conclude the proof of the testability of linear transformations (Theorem 1.1). Our presentation of the BSG theorem in Section 3 follows the one by Sudakov, Szemerédi, and Vu [18], which relies on a graph-theoretic lemma regarding certain paths in dense graphs. In Section 6 we also present the proof of the optimality of the path length of this lemma, due to Kostochka and Sudakov [14]. This section is not needed for the proof of Theorem 1.1; we include it to provide a more complete picture of the proof techniques we present. For the same reason, in Section 7 we present a simpler proof of the testability of linear transformations in the case in which the agreement is large (corresponding to $\varepsilon \approx 1$).

2 Preliminaries

In this work we are concerned with subsets of the vector space $\mathbb{F}_2^n$, whose operation is the componentwise addition mod 2 denoted by "+." Throughout this survey, A denotes a subset of $\mathbb{F}_2^n$.

For sets $A, B \subseteq \mathbb{F}_2^n$, we denote by $A+B$ the set $\{a+b : a \in A, b \in B\}$. For an integer ℓ we denote by ℓA the set $A+A+\cdots+A$ where the number of summands is ℓ. Finally, we denote by $\mathrm{span}(A)$ the span of the elements of A, i. e., $\mathrm{span}(A) = \bigcup_\ell \ell A$.

We use several times the following basic counting argument, whose proof is straightforward.

Proposition 2.1. *Let $f : D \to S$ be a function, for finite sets D and S. If it holds that $|f^{-1}(s)| \geq t$ for every $s \in S$, then $|S| \leq |D|/t$.*

Finally, all the graphs in this paper are undirected and have no self-loops.

3 The Balog-Szemerédi-Gowers (BSG) theorem

In this section we prove the Balog-Szemerédi-Gowers (BSG) theorem [2, 9], which is stated next.

Theorem 3.1 (Balog-Szemerédi-Gowers)**.** *For all $\varepsilon > 0$, for all sufficiently large n and all sets $A \subseteq \mathbb{F}_2^n$, the following holds. If $\Pr_{a,a' \in A}[a+a' \in A] \geq \varepsilon$ then there is $A' \subseteq A$, with $|A'| \geq (\varepsilon/3) \cdot |A|$, such that $|2A'| \leq (6/\varepsilon)^8 \cdot |A|$.*

One way to think of the BSG theorem is the following. For a subset E of the cartesian product $A \times A$, let us denote its set of sums by $\sum E := \{a+b : (a,b) \in E\}$. Then the BSG theorem says that from a dense $E \subseteq A \times A$ such that $\sum E$ is a subset of A and hence is small compared to $|E|$, we can obtain a dense $A' \subseteq A$ such that $|\sum(A' \times A')| = |2A'|$ is small compared to $|A'|^2$.

The proof that we present of the above Theorem 3.1 follows one due to Sudakov, Szemerédi, and Vu [18]. It makes use of the following graph-theoretical statement, which does not use any property of addition and only relies on the density of the graph. The use of the language of graph theory to prove results in additive combinatorics has been found fruitful, and it goes back at least to Szemerédi's theorem on arithmetic progressions [19].

Lemma 3.2 (Sudakov-Szemerédi-Vu)**.** *Let $G = (A, E)$ be a graph with $|A| = N$ nodes, $|E| = \varepsilon \cdot N^2$ edges. Then there is a set $A' \subseteq A$, $|A'| \geq \varepsilon \cdot N$ such that for every $a, b \in A'$ there are at least $(\varepsilon/2)^8 \cdot N^3$ paths of length 4 in G from a to b.*

Proof of Theorem 3.1, assuming Lemma 3.2. Consider the graph $G = (A, E)$ on $|A|$ nodes where two distinct nodes are adjacent if and only if their sum is in A, i. e., $E := \{\{a,b\} : a+b \in A, a \neq b\}$. By assumption, $|E| \geq (\varepsilon/3) \cdot |A|^2$. (For large values of n, the factor $1/3$ generously accounts for the translation between the hypothesis, which talks about pairs, and Lemma 3.2, which talks about edges.) Now let A' be the subset of A given by Lemma 3.2, and consider any two $a, b \in A'$. By Lemma 3.2, there are $\varepsilon' \cdot |A|^3$ paths (a, c_1, c_2, c_3, b) with edges in E, where $\varepsilon' = (\varepsilon/6)^8$. By the definition of E, the sum of two consecutive nodes in any path lies in A. Thus, considering the function $f(x_1, x_2, x_3, x_4) := x_1 + x_2 + x_3 + x_4$, we have that, for every $a, b \in A'$,

$$f(a+c_1, c_1+c_2, c_2+c_3, c_3+b) = (a+c_1)+(c_1+c_2)+(c_2+c_3)+(c_3+b) = a+b\,,$$

for at least $\varepsilon' \cdot |A|^3$ inputs

$$(x_1, x_2, x_3, x_4) := (a+c_1, c_1+c_2, c_2+c_3, c_3+b) \in A^4.$$

Note that distinct triples (c_1,c_2,c_3) give rise to distinct inputs (x_1,x_2,x_3,x_4), which follows from the fact that a and b are fixed. In other words, we have shown that each element of $2A'$ can be represented in at least $\varepsilon'|A|^3$ different ways as a sum of 4 elements of A. Via Proposition 2.1 this leads to the following upper bound on $|2A'|$:

$$|2A'| \leq |A|^4/(\varepsilon' \cdot |A|^3)\,,$$

concluding the proof. □

Proof of Lemma 3.2. The idea is to exhibit a set $A' \subseteq A$ such that every $a \in A'$ shares $\Omega(N)$ neighbors with most nodes in A'. (We may think of "most" as a 0.9 fraction.) From this we infer that, for every two nodes $a,b \in A'$, most nodes c_2 in A' share $\Omega(N)$ neighbors c_1 with a and also share $\Omega(N)$ neighbors c_3 with b, which implies the result. We now give the details.

For a node $v \in G$ let us denote by $N(v) \subseteq A$ the neighborhood of v. The set A' will be a subset of $N(v')$ for some v' given by a probabilistic argument. For this argument, let us call a pair $\{u,w\}$ of distinct vertices *bad* if $|N(u) \cap N(w)| \leq \varepsilon^3 \cdot N$. Let, moreover, $v \in G$ be a uniformly distributed random node of G. We are interested in the number of bad pairs inside $N(v)$. Let $B_{\{u,w\}}$ be the $0/1$ indicator variable that is 1 when $u,v \in N(v)$ and $\{u,w\}$ is bad. For every bad pair $\{u,w\}$ (not necessarily in $N(v)$) it holds that $\{u,w\} \subseteq N(v)$ if and only if v is a common neighbor of u and w, which by the definition of "bad" happens with probability at most ε^3. Consequently, by the linearity of expectation, we have

$$E_{v \in A}[\text{number of bad pairs in } N(v)] \leq \varepsilon^3 \cdot \binom{N}{2} \leq \varepsilon^3 \cdot N^2/2\,. \tag{3.1}$$

Let us now denote by $S(v)$ the set of nodes $u \in N(v)$ that form a bad pair with at least $\varepsilon^2 \cdot N$ other nodes $w \in N(v)$. Since there are always at least $|S(v)| \cdot \varepsilon^2 \cdot N/2$ bad pairs in $N(v)$, where the factor $1/2$ comes from the fact that each bad $\{u,w\}$ is counted once for u and once for w, Equation (3.1) implies that

$$E_{v \in A}[|S(v)|] \leq (\varepsilon^3 \cdot N^2/2)/(\varepsilon^2 \cdot N/2) = \varepsilon \cdot N\,. \tag{3.2}$$

Therefore, using the fact that $E[|N(v)|] = 2 \cdot \varepsilon \cdot N$ because the graph has $\varepsilon \cdot N^2$ edges, we have

$$E_{v \in A}[|N(v) \setminus S(v)|] = E[|N(v)|] - E[|S(v)|] \geq 2 \cdot \varepsilon \cdot N - \varepsilon \cdot N = \varepsilon \cdot N\,.$$

We now fix a $v' = v$ that maximizes the quantity $|N(v) \setminus S(v)|$ and let $A' := N(v') \setminus S(v')$ be the corresponding set; therefore $|A'| \geq \varepsilon \cdot N$.

To see that A' satisfies the conclusion of the lemma, consider any $a,b \in A'$. Since we removed the nodes in $S(v')$, i. e., those that form a bad set with at least $\varepsilon^2 \cdot N$ other nodes $w \in N(v')$, both a and b form a good pair with all but at most $\varepsilon^2 \cdot N$ nodes of A'. So there are at least

$$|A'| - 2 \cdot \varepsilon^2 \cdot N \geq \varepsilon \cdot N - 2 \cdot \varepsilon^2 \cdot N \geq \varepsilon^2 \cdot N$$

nodes $c_2 \in A'$ that form a good set with both a and b, where the last inequality holds if we assume that $\varepsilon \leq 1/3$. (To optimize exponents, one can replace the last inequality with $\varepsilon - 2\varepsilon^2 \geq \varepsilon/3$.) For every such c_2 we have, by the definition of "good," $\varepsilon^3 \cdot N$ choices for c_1 and $\varepsilon^3 \cdot N$ choices for c_3 such that (a,c_1,c_2,c_3,b) is a path in G. In total, we have at least $(\varepsilon^3 \cdot N)(\varepsilon^2 \cdot N)(\varepsilon^3 \cdot N) = \varepsilon^8 \cdot N^3$ such paths in G. This proves the theorem under the the assumption that $\varepsilon \leq 1/3$. If $\varepsilon > 1/3$, the same proofs works when ε is replaced with $\varepsilon/2$, which is at most $1/3$ because any graph trivially has at most $N^2/2 \geq \varepsilon \cdot N^2$ edges. □

4 Ruzsa's theorem

In this section we prove Ruzsa's theorem [16], which states that the span of A does not expand too much if $2A$ does not.

Theorem 4.1 (Ruzsa)**.** *For all c there is c' such that for all n and all sets $A \subseteq \mathbb{F}_2^n$ the following holds. If $|2A| \le c \cdot |A|$ then $|\operatorname{span}(A)| \le c' \cdot |A|$.*

The core of the proof of Theorem 4.1 is the following lemma, which states that $4A$ does not expand too much if $2A$ does not.

Lemma 4.2. *For all n and all sets $A \subseteq \mathbb{F}_2^n$ we have $|4A| \le 16 \cdot (|2A|/|A|)^4 \cdot |2A|$.*

Proof of Theorem 4.1 assuming Lemma 4.2. We start with the following *covering claim* showing that we can cover all of $4A$ by few translates of A: There is a set $X \subseteq 3A$ whose size depends only on c such that for every $b \in 3A$ we have

$$|(X+A)\cap(b+A)| \ge 1\,. \tag{4.1}$$

To prove the covering claim, initialize X to the empty set, and as long as there is some $b \in 3A$ violating (4.1), add b to X. The resulting X satisfies the intersection requirement by construction. To verify the bound on the size of X, note that at each iteration the set $X+A$ grows in size by $|A|$, but $X+A \subseteq 4A$ always holds, and so at the end of the process $|X|$ is at most $|4A|/|A|$. By Lemma 4.2, together with our assumption that $|2A| \le c\cdot|A|$, it follows that this quantity depends only on c.

Now we show by induction that, for every $\ell \ge 3$, $\ell A \subseteq (\ell-2)X + 2A$. This will conclude the proof, as the size of X depends only on c. Specifically we obtain that $\operatorname{span}(A) \subseteq \operatorname{span}(X)+2A$, and so $|\operatorname{span}(A)| \le |\operatorname{span}(X)|\cdot|2A| \le 2^{|X|}\cdot c\cdot|A|$.

For the base case $\ell = 3$ of the induction, take any $b \in 3A$. By (4.1), $X+A$ intersects $b+A$, which means $x+a = b+a'$ for some $x \in X$ and $a,a' \in A$, and so $b = x+a+a' \in X+2A$.

For the inductive step, write

$$\ell A = (\ell-1)A + A \subseteq (\ell-3)X + 2A + A \subseteq (\ell-2)X + 2A\,,$$

where we apply the inductive hypothesis and then the base case. □

Proof of Lemma 4.2. We start with the following *covering claim*, whose statement and proof are very similar to those of the covering claim in the proof of Theorem 4.1 above. (We omit the proof.) For any $A \subseteq \mathbb{F}_2^n$ there is a set $X \subseteq A$ of size $|X| \le 2\cdot|2A|/|A|$ such that for every $b \in A$ we have

$$|(X+A)\cap(b+A)| \ge |A|/2\,.$$

As a consequence of the covering claim, we have that for every $b \in A$ there are at least $|A|/2$ triples $(a_0,a_1,x) \in A\times A\times X$ such that $b = x+a_0+a_1$. This is because each element $y \in (X+A)\cap(b+A)$ gives rise to, say, one such triple with $a_1 := b+y$. (This last requirement makes all the triples distinct.) Now we use this implication to prove the lemma. The idea is to represent each element w of $4A$ in at least $(|A|/2)^2$ distinct ways as a sum of a quintuple (c,c',c'',x,x') with $c,c',c'' \in 2A$ and $x,x' \in X$; then an application of Proposition 2.1 completes the proof. More specifically, write $w \in 4A$ as $w = z+z'$ for $z,z' \in 2A$. We first represent each of z and z' separately in many ways as a sum of a triple, then we represent the pair (z,z') in many ways as a sum of a sextuple, and finally we map the sextuples in a one-to-one fashion into quintuples that represent the sum $z+z'=w$ in many ways. We now give the details.

Fix an arbitrary $w \in 4A$, and fix some $z,z' \in 2A$ for which $w=z+z'$. Further write $z=b_0+b_1$, for $b_0,b_1 \in A$. By the above, there are at least $|A|/2$ triples $(a_0,a_1,x) \in A \times A \times X$ such that $z=b_0+a_0+a_1+x$. Since $b_0+a_0 \in 2A$, there are at least $|A|/2$ triples $(c,a_1,x) \in (2A) \times A \times X$ such that $z=c+a_1+x$. By repeating the argument for $z'=b'_0+b'_1$, we obtain that there are at least $(|A|/2)^2$ sextuples

$$(c,c',a_1,a'_1,x,x') \in (2A) \times (2A) \times A \times A \times X \times X$$

such that

$$\begin{aligned} z &= c+a_1+x\,, \\ z' &= c'+a'_1+x'\,. \end{aligned}$$

Note that, in any solution to the above system, a_1 and a'_1 are uniquely determined once c,x,c',x' are chosen (as z and z' are fixed). So, two different solutions cannot differ only in the two coordinates ranging in A. In particular, the map that takes a solution

$$(c,c',a_1,a'_1,x,x') \in (2A) \times (2A) \times A \times A \times X \times X$$

to the quintuple

$$(c,c',a_1+a'_1,x,x') \in (2A) \times (2A) \times (2A) \times X \times X$$

is one-to-one. Moreover, such a quintuple sums up to $z+z'=w$.

Therefore, similar to the proof of Theorem 3.1, we have a function

$$f(x_1,x_2,x_3,x_4,x_5) := x_1+x_2+x_3+x_4+x_5$$

such that, for every element $w \in 4A$, there are at least $(|A|/2)^2$ distinct inputs y such that $f(y)=w$. By Proposition 2.1, we have

$$|4A| \le |2A|^3 \cdot |X|^2 \cdot 4/|A|^2 \le 16 \cdot (|2A|/|A|)^4 \cdot |2A|\,,$$

where we are using the fact, established at the beginning of this proof, that $|X| \le 2 \cdot |2A|/|A|$. □

Remark 4.3 (On the loss in parameters)**.** We remark that one can eliminate the factor 16 in Lemma 4.2 by applying the lemma as stated to the set $A \times A \times \cdots \times A$. (See [20, Corollary 2.18], for example.) Turning back to the main result of this section, Theorem 4.1, we note that the current best upper bound on c' is obtained by Green and Tao [11] who prove $c' \le 2^{2c}$ modulo lower order factors. Taking A to be a set of $2c$ independent vectors one notes that $c' \le 2^{2c}$ is the best possible, and in particular that c' in general must be exponential in c.

However, if one is willing to settle for the span of a large subset A' of A, rather than all of A, in the same spirit as the BSG theorem, then it is conjectured [10] that c' can be made polynomial in c. This would imply a polynomial dependence of ε' on ε in Theorem 1.1.

5 Obtaining a linear transformation

In this section we conclude the proof of the property testing result in Theorem 1.1. The last component of the proof is the following linear-algebraic fact that states that if the span of (a large subset of) $\{(x, f(x)) : x \in \mathbb{F}_2^n\}$ does not grow much, then f is approximately a linear transformation.

Lemma 5.1. *For all $\varepsilon > 0$, for all sufficiently large n, all functions $f : \mathbb{F}_2^n \to \mathbb{F}_2^n$, and all sets $A \subseteq \{(x, f(x)) : x \in \mathbb{F}_2^n\} \subseteq \mathbb{F}_2^{2n}$, the following holds. If*

$$\varepsilon \cdot 2^n \leq |A| \leq |\operatorname{span}(A)| \leq 2^n/\varepsilon\,,$$

then there is an $n \times n$ matrix M such that

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) = Mx] \geq \varepsilon^3/3\,.$$

Proof. We start by finding an affine transformation $Tx + u$ with the required property, then we observe how this implies the existence of a linear transformation. Let $v_1, v_2, \ldots, v_a$ be a basis of $\operatorname{span}(A)$. By definition, every vector $(x, f(x)) \in A$ is a linear combination of the v_i, i. e., for every $(x, f(x)) \in A$ there exists $w \in \mathbb{F}_2^a$ such that, in matrix form,

$$\begin{array}{|c|}\hline x \\ \hline f(x) \\ \hline \end{array} = \begin{array}{|cccc|}\hline | & | & & | \\ v_1 & v_2 & \cdots & v_a \\ | & | & & | \\ \hline \end{array} \cdot w\,.$$

Let us now add to our collection new vectors $v_{a+1}, v_{a+2}, \ldots, v_k$ so that the projection onto the first n coordinates of $\operatorname{span}(\{v_1, \ldots, v_k\})$ is all of $\mathbb{F}_2^n$. In order to bound k, note that since $A \subseteq \{(x, f(x)) : x \in \mathbb{F}_2^n\}$, the projection of A on the first n coordinates has size $\geq \varepsilon \cdot 2^n$. Iteratively adding vectors that double this projection, we see that we can choose $k - a \leq \log(1/\varepsilon)$. Also note that $a \leq n + \log(1/\varepsilon)$ by assumption, hence $k \leq n + 2\log(1/\varepsilon)$. Let V_k be the resulting matrix of the vectors $v_1, \ldots, v_k$. By performing Gaussian elimination on the columns, we can find an invertible transformation that brings V_k into the following canonical form:

$$\begin{array}{|c|c|}\hline I & 0 \\ \hline T & U \\ \hline \end{array}\,,$$

where I is the $n \times n$ identity matrix, T is also $n \times n$, and U is $n \times (k - n)$. This is possible because the projection of the vectors $v_1, \ldots, v_k$ onto the first n coordinates spans $\mathbb{F}_2^n$. In other words, there is an invertible $k \times k$ matrix L such that $V_k \cdot L$ is in the canonical form. Since L is invertible, we still have the property that for every vector $(x, f(x)) \in A$ there exists $w \in \mathbb{F}_2^k$ such that

$$\begin{array}{|c|}\hline x \\ \hline f(x) \\ \hline\end{array} = \begin{array}{|c|c|}\hline I & 0 \\ \hline T & U \\ \hline\end{array} \cdot w\,.$$

This means that the first n coordinates of w must equal x. Consequently, for every $(x, f(x)) \in A$ it holds $f(x) = Tx + Uz$, for some z of length $k - n \leq 2\log(1/\varepsilon)$. Therefore, by an averaging argument, there exists a fixed $u = Uz$ so that

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) = Tx + u] \geq \varepsilon \cdot 2^{-(k-n)} \geq \varepsilon^3\,,$$

where we also use the fact that the projection of A on the first n coordinates has size $\geq \varepsilon \cdot 2^n$.

This gives us an affine transformation, and in what follows we show how one can get a linear transformation, i. e., get rid of the 'u' above, with only a slight loss in probability. We claim that

$$(\exists i \leq n)\left[\Pr_{x \in \mathbb{F}_2^n}[f(x) = Tx + u \,|\, x_i = 1] \geq 0.99 \cdot \varepsilon^3\right]. \tag{5.1}$$

Such a claim lets us construct a linear transformation M by summing u to the i-th column of T (in other words, $Mx = Tx + x_i \cdot u$), concluding the proof of the lemma. The different factor in the conclusion of the lemma generously accounts for the probability that $x_i = 1$.

It remains to prove (5.1). For this, let X be the uniform distribution over $\mathbb{F}_2^n$ and let Y be the distribution on $\mathbb{F}_2^n$ that is obtained by selecting a random index $i \leq n$, setting to 1 the i-th bit, and choosing the other bits uniformly at random. We would like to argue that there is not much difference in working with X or Y. This is useful because if we work with Y we easily obtain (5.1) by an appropriate choice of the index i in the definition of Y. To formalize this, consider the statistical distance between X and Y, i. e., the maximum over all sets S of $|\Pr[X \in S] - \Pr[Y \in S]|$. (While we are interested in $S := \{x : f(x) = Tx + u\}$, the following applies to any S.)

Note that this distance is maximized by the set $\bar{S}$ of strings of weight at most $n/2$, which is the set of strings having larger probability according to X than Y. Also, by Stirling's approximation [6, Lemma 17.5.1], both $\Pr[X \in \bar{S}]$ and $\Pr[Y \in \bar{S}]$ are $1/2 + \Theta(1/\sqrt{n})$. Therefore for large n their distance is at most $0.01 \cdot \varepsilon^3$, and in particular

$$\Pr_{x \in Y}[f(x) = Tx + u] \geq 0.99 \cdot \varepsilon^3\,,$$

which proves (5.1). □

We can now paste everything together for a quick conclusion of the proof of Theorem 1.1 about testability of linear transformations.

Proof of Theorem 1.1. The proof amounts to defining $A := \{(x, f(x)) : x \in \mathbb{F}_2^n\} \subseteq \mathbb{F}_2^{2n}$ and composing Theorems 3.1 and 4.1, and Lemma 5.1. □

6 Optimality of the path length in Lemma 3.2

In this section we discuss the optimality of the path length in the graph-theoretic Lemma 3.2 that is the core of the proof of the BSG theorem. Recall that the lemma establishes that every dense graph contains a large subset of nodes such that every two nodes in the subset are connected by many paths of length 4. It is natural to ask if the path length can be reduced from 4, and one can quickly see that it cannot be set to 3: the graph could be bipartite, for instance, and any set of at least 3 nodes would have two nodes on the same side which cannot be connected by any path of odd length 3. We now state and prove a result by Kostochka and Sudakov [14] that also rules out path length 2. Thus, path length 4 is optimal in Lemma 3.2.

Theorem 6.1 (Kostochka-Sudakov)**.** *For all $\varepsilon > 0$ there exist arbitrarily large values of N such that there is a graph on N vertices with $N^2/4$ edges such that in every set of $\varepsilon \cdot N$ nodes there are two nodes with less than $\varepsilon \cdot N$ common neighbors.*

In fact, this is true for all sufficiently large values of N but we only prove it for certain powers of 2.

6.1 Proof of Theorem 6.1

Let n be a sufficiently large even integer, and let $N := 2^n$. Identify the set of N nodes with the binary strings of length n. Let $\Delta(u,v)$ denote the Hamming distance between nodes u and v, i. e., the number of positions i such that $u_i \neq v_i$, and connect two nodes $u \neq v$ if and only if $\Delta(u,v) \leq n/2$. Each node has at least $N/2$ neighbors, and thus the graph has at least $N(N/2)/2 = N^2/4$ edges. We now show that it also has the desired property. The main idea is that any set of $\varepsilon \cdot N$ nodes must contain two nodes at Hamming distance at least $n - O(\sqrt{n})$, but two such nodes have less than $\varepsilon \cdot N$ common neighbors.

We now present the formal proof, starting with the next claim that gives us two distant nodes.

Claim 6.2. *Let S be any set of $\varepsilon \cdot N$ nodes. Then S contains two nodes at Hamming distance at least $n - c \cdot \sqrt{n}$, where c is a constant that only depends on ε.*

Let us give some details on how the claim is proved. (For a somewhat different argument, see [22].)

The proof will use the following lemma from a set of notes by Barvinok. The history of this lemma goes back to Harper [13], and similar lemmas can be found elsewhere, see for example [15, Theorem 14.2.3]. The following formulation is particularly useful to us because it is not limited to sets of measure 1/2.

Lemma 6.3 (Corollary 4.4 in [3])**.** *Let $S \subseteq \mathbb{F}_2^n$ be a non-empty set. Then, for any $b > 0$, we have*

$$\frac{|\{x \in \mathbb{F}_2^n : \forall y \in S \;\; \Delta(x,y) \geq b\sqrt{n}\}|}{2^n} \leq \frac{2^n}{|S|} e^{-b^2} .$$

Proof sketch of Claim 6.2. First, note that if $\varepsilon > 1/2$ then the claim is easily proved with maximal Hamming distance n, that is, we can find two nodes u and v such that $\Delta(u,v) = n$. This is because we can pair off each node with its complement at distance n, and a set S of size $|S| \geq \varepsilon \cdot N > N/2$ must take both nodes from some pair.

To handle the case $\varepsilon \leq 1/2$, we apply the argument to the set S' of nodes u such that there exists $v \in S$ at distance $\Delta(u,v) \leq b \cdot \sqrt{n}$. Specifically, by applying Lemma 6.3 with a large enough constant b depending only on ε, we obtain that this set S' has size $> 2^n/2$. We can now apply the previous "$\varepsilon > 1/2$" argument to S', from which the claim follows with $c = 2 \cdot b$. □

Now that we have these two nodes at distance $n - c \cdot \sqrt{n}$, we conclude the proof of Theorem 6.1 by showing that the number of their common neighbors is less than $\varepsilon \cdot N$. Without loss of generality, let these two nodes, which we denote by u_1 and u_2, be the all-zero vector and the vector that is 0 exactly in the first $k := c \cdot \sqrt{n}$ coordinates, respectively. Let us now see what nodes are common neighbors of u_1 and u_2. Let $X \in \mathbb{F}_2^n$ be a node, and let $P = P(X)$ be its number of 1's in the first k coordinates, and $Q = Q(X)$ its number of 1's in the other $n-k$ coordinates. The node X is a common neighbor of u_1 and u_2 precisely when $P + Q \leq n/2$ and $P + (n-k-Q) \leq n/2$. By combining the inequalities, we obtain that

$$n/2 - k + P \leq Q \leq n/2 - P\,, \tag{6.1}$$

i. e., for a given P, if X is a neighbor of both u_1 and u_2 then Q has to lie in a set of $k - 2 \cdot P + 1$ integers.

The intuition for the rest of the proof is as follows. A typical P is within $O(\sqrt{k})$ of $k/2$, and by (6.1) such a P constricts Q to lie in a set of $k - 2 \cdot P + 1 = O(\sqrt{k})$ integers. As is well known, by Stirling's approximation [6, Lemma 17.5.1] the probability that Q is equal to any particular integer is $O(1/\sqrt{n-k})$. Since $k = O(\sqrt{n})$ and n is large, we have $O(1/\sqrt{n-k}) = o(1/\sqrt{k})$, and so by a union bound Q falls in the set of $O(\sqrt{k})$ integers with probability tending to 0, and this proves the theorem.

More formally, let $d = d(\varepsilon)$ be a sufficiently large constant to be determined later. Let us choose a random node X, and let $P = P(X)$ and $Q = Q(X)$ respectively denote the Hamming weight of its first k and last $n-k$ bits. We have:

$$\begin{aligned}
\Pr_{P,Q}[\,(6.1) \text{ holds}\,] &\leq \Pr_Q\left[(6.1) \text{ holds} \,\middle|\, |P - k/2| \leq d \cdot \sqrt{k}\right] + \Pr_P\left[|P - k/2| > d \cdot \sqrt{k}\right] \\
&\leq 2 \cdot d \cdot \sqrt{k} \cdot \Pr_Q[Q = (n-k)/2] + \varepsilon/2 \\
&\leq O(d \cdot \sqrt{k}/\sqrt{n-k}) + \varepsilon/2 \\
&< \varepsilon\,.
\end{aligned}$$

To bound the term

$$\Pr_Q\left[(6.1) \text{ holds} \,\middle|\, |P - k/2| \leq d \cdot \sqrt{k}\right]$$

we use a union bound (where actually a factor 2 could be saved noting that if $P > k/2$ then (6.1) does not hold) and the fact that $\Pr_Q[Q = (n-k)/2] \geq \Pr_Q[Q = (n-k)/2 + r]$ for any $r \in \mathbb{R}$. To bound the term

$$\Pr_P\left[|P - k/2| > d \cdot \sqrt{k}\right]$$

we apply a Chernoff Bound (see, e. g., [7]), choosing $d = d(\varepsilon)$ to be sufficiently large. Later, we use Stirling's approximation [6, Lemma 17.5.1] to bound $\Pr_Q[Q = (n-k)/2]$. Finally, the last inequality holds for sufficiently large n recalling that $k = c \cdot \sqrt{n}$.

7 Testing linear transformations when the agreement is large

In this section we show a simpler proof of Theorem 1.1 in the case in which the agreement is large. This proof was communicated to us by Shachar Lovett; it uses the ideas in [1].

Theorem 7.1. *For all $\gamma \in [0,1/8)$, all n, and all functions $f : \mathbb{F}_2^n \to \mathbb{F}_2^n$, the following holds. If*

$$\Pr_{x,x' \in \mathbb{F}_2^n}[f(x) + f(x') = f(x+x')] \geq 1 - \gamma,$$

then there is an $n \times n$ matrix M such that

$$\Pr_{x \in \mathbb{F}_2^n}[f(x) = Mx] \geq 1 - 2\gamma.$$

The constant 2 in Thm 7.1 can be somewhat reduced at the cost of a more complicated proof.

7.1 Proof of Theorem 7.1

Define $g(x)$ to be the "majority vote," i. e., a value that maximizes $\Pr_y[f(x+y)+f(y) = g(x)]$. First, we claim that

$$\Pr_x[f(x) \neq g(x)] \leq 2 \cdot \gamma.$$

To see this, note that

$$\Pr_x[\Pr_y[f(x) \neq f(y) + f(x+y)] \geq 1/2] \leq 2 \cdot \gamma,$$

for else the assumption is contradicted. So, for all but a $2 \cdot \gamma$ fraction of x, we have that

$$\Pr_y[f(x+y) + f(y) = f(x)] > \frac{1}{2}$$

and consequently $g(x) = f(x)$.

To complete the proof, it remains to see that g is linear. The first step is to show that the majority in the definition of g is always overwhelming.

Claim 7.2. *For every x, $\Pr_y[g(x) \neq f(x+y) + f(y)] \leq 2 \cdot \gamma$.*

Proof. By the union bound, we have

$$\begin{aligned}&\Pr_{y,z}[f(x+y) + f(y) \neq f(x+z) + f(z)] \\ &\qquad\leq \Pr_{y,z}[f(y) + f(z) \neq f(y+z)] + \Pr_{y,z}[f(x+y) + f(x+z) \neq f(y+z)] \leq 2 \cdot \gamma.\end{aligned}$$

So there is a fixed z such that $\Pr_y[f(x+y) + f(y) \neq f(x+z) + f(z)] \leq 2 \cdot \gamma$. Since $\gamma < 1/4$, $g(x) = f(x+z) + f(z)$. □

To see that g is linear, fix any x, x', choose y, y' uniformly at random, and consider the following 5 equations (3 rows and 2 columns):

$$\begin{array}{ccccc} g(x) & = & f(x+y) & + & f(y) \\ & & + & & + \\ g(x') & = & f(x'+y') & + & f(y') \\ & & = & & = \\ g(x+x') & = & f(x+x'+y+y') & + & f(y+y')\,. \end{array}$$

By applying Claim 7.2 to the rows, and the theorem's hypothesis to the columns, and using a union bound, we see that with probability at least $1-3\cdot 2\cdot\gamma-2\cdot\gamma=1-8\cdot\gamma>0$ all the 5 equations hold, which means that $g(x)+g(x')=g(x+x')$. Since x and x' were arbitrary, g is linear and one can write $g(x)=Mx$ for an $n\times n$ matrix M.

Acknowledgment We thank Andrej Bogdanov, Vladimir Trifonov, and Avi Wigderson for helpful discussions, Benny Sudakov for pointing out and explaining to us the paper [14], and Shachar Lovett for telling us the proof of Theorem 7.1. We also thank the Columbia theory reading group for the opportunity to present this material [22] in Fall 2007 and for helpful comments. Finally, we are very grateful to Oded Regev and the anonymous referees for extensive feedback.

References

[1] NOGA ALON, TALI KAUFMAN, MICHAEL KRIVELEVICH, SIMON LITSYN, AND DANA RON: Testing low-degree polynomials over GF(2). In *Proc. 7th Intern. Workshop Randomization Approx. Tech. in Comput. Sci. (RANDOM)*, volume 2764 of *Lecture Notes in Comput. Sci.*, pp. 188–199. Springer, 2003. [doi:10.1007/978-3-540-45198-3_17] 12

[2] ANTAL BALOG AND ENDRE SZEMERÉDI: A statistical theorem of set addition. *Combinatorica*, 14(3):263–268, 1994. [doi:10.1007/BF01212974] 2, 4

[3] ALEXANDER BARVINOK: Measure Concentration, 2005. Lecture notes. Available at `http://www.math.lsa.umich.edu/~barvinok/total710.pdf`. 10

[4] MIHIR BELLARE, DON COPPERSMITH, JOHAN HÅSTAD, MARCOS A. KIWI, AND MADHU SUDAN: Linearity testing in characteristic two. *IEEE Trans. Inform. Theory*, 42(6):1781–1795, 1996. [doi:10.1109/18.556674] 2

[5] MANUEL BLUM, MICHAEL LUBY, AND RONITT RUBINFELD: Self-testing/correcting with applications to numerical problems. *J. Comput. System Sci.*, 47(3):549–595, 1993. [doi:10.1016/0022-0000(93)90044-W] 2

[6] THOMAS M. COVER AND JOY A. THOMAS: *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006. 9, 11

[7] DEVDATT P. DUBHASHI AND ALESSANDRO PANCONESI: *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009. 11

[8] GREGORY A. FREĬMAN: *Foundations of a Structural Theory of Set Addition*. American Mathematical Society, 1973. Translated from the Russian, Translations of Mathematical Monographs, Vol 37. 3

[9] TIMOTHY GOWERS: A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geom. Funct. Anal.*, 8(3):529–551, 1998. [doi:10.1007/s000390050065] 2, 4

[10] BEN GREEN: Finite field models in additive combinatorics. In *Surveys in Combinatorics*, number 327 in London Math. Soc. Lecture Note Ser., pp. 1–28. Cambridge University Press, 2005. [doi:10.1017/CBO9780511734885.002] 2, 7

[11] BEN GREEN AND TERENCE TAO: Freiman's theorem in finite fields via extremal set theory. *Combin. Probab. Comput.*, 18(3):335–355, 2009. [doi:10.1017/S0963548309009821] 7

[12] BEN GREEN AND TERENCE TAO: A note on the Freiman and Balog-Szemerédi-Gowers theorems in finite fields. *J. Aust. Math. Soc.*, 86(1):61–74, 2009. [doi:10.1017/S1446788708000359] 3

[13] L. H. HARPER: Optimal assignments of numbers to vertices. *SIAM Journal on Applied Mathematics*, 12(1):131–135, 1964. 10

[14] ALEXANDR KOSTOCHKA AND BENNY SUDAKOV: On Ramsey numbers of sparse graphs. *Combin. Probab. Comput.*, 12(5-6):627–641, 2003. [doi:10.1017/S0963548303005728] 3, 10, 13

[15] JIŘÍ MATOUŠEK: *Lectures on Discrete Geometry*. Springer-Verlag, 2002. 10

[16] IMRE Z. RUZSA: An analog of Freiman's theorem in groups. *Astérisque*, 258:xv, 323–326, 1999. 3, 6

[17] ALEX SAMORODNITSKY: Low-degree tests at large distances. In *Proc. 39th STOC*, pp. 506–515. ACM Press, 2007. [doi:10.1145/1250790.1250864] 1, 3

[18] BENNY SUDAKOV, ENDRE SZEMERÉDI, AND VAN VU: On a question of Erdős and Moser. *Duke Math. J.*, 129(1):129–155, 2005. [doi:10.1215/S0012-7094-04-12915-X] 3, 4

[19] ENDRE SZEMERÉDI: On sets of integers containing no k elements in arithmetic progression. *Acta Arith.*, 27:199–245, 1975. [doi:10.1007/BF01894569] 4

[20] TERENCE TAO AND VAN H. VU: *Additive Combinatorics*. Volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2006. 1, 3, 7

[21] LUCA TREVISAN: Additive combinatorics and theoretical computer science. *SIGACT News*, 40(2):50–66, 2009. [doi:10.1145/1556154.1556170] 1

[22] EMANUELE VIOLA: Selected results in additive combinatorics: An exposition. Technical Report 103, Electron. Colloq. on Comput. Complexity (ECCC), 2007. `http://www.eccc.uni-trier.de/report/2007/103/`. 1, 10, 13

AUTHOR

Emanuele Viola
Assistant Professor
Northeastern University, Boston, MA
viola@ccs.neu.edu
http://www.ccs.neu.edu/home/viola/

ABOUT THE AUTHOR

EMANUELE VIOLA graduated from Harvard University in 2006 under the supervision of Salil Vadhan. He wrote most of this survey in 2007, while he was a postdoctoral fellow in Avi Wigderson's group at the Institute for Advanced Study in Princeton, NJ. This year he is trying to learn to play squash.