Marton's Polynomial Freiman-Ruzsa conjecture

Timothy Gowers, Ben Green, Freddie Manners, Terence Tao

University of California, Los Angeles

Inverse sumset theorems

- A key foundational topic in modern additive combinatorics is that of inverse theorems - theorems that show that objects with large amounts of "approximate additive structure" must in fact be close to objects with "exact additive structure".
- An influential early example of an inverse theorem (stated in modern language) is

Freiman's theorem (1964)

If $A \subset \mathbf{Z}$ is finite non-empty with doubling constant at most K, then A is contained in a convex progression P of cardinality at most $f(K)|A|$ and rank at most $d(K)$, for some functions f, d.

- Here, the doubling constant of A is the quantity $K := |A + A|/|A|$, where $A \pm B := \{a \pm b : a \in A, b \in B\}$ denotes the sumset or difference set of A and B, and $|A|$ denotes the cardinality of A.
- Note that $K \geq 1$. We will be interested primarily in the regime where K is somewhat large (and $|A|$ even larger).
- An convex progression of rank d in an additive group G is a set of the form

$$\{n_1 v_1 + \cdots + n_d v_d : (n_1, \ldots, n_d) \in B \cap \mathbf{Z}^d\}$$

for some symmetric convex body $B \subset \mathbf{R}^d$ and $v_1, \ldots, v_d \in G$.

- This is a modern version of the notion of a generalized arithmetic progression.

There is an equivalent form of Freiman's theorem, in which containment $A \subset P$ is replaced by covering $A \subset P + S$:

Freiman's theorem, alternate form

If $A \subset \mathbf{Z}$ is finite non-empty with doubling constant at most K, then A can be covered by at most $g(K)$ translates of a convex progression P of cardinality at most $f(K)|A|$ and rank at most $d(K)$, for some functions f, g, d.

While seemingly weaker, this form has more efficient quantitative dependencies on K.

Freiman's theorem, alternate form

If $A \subset \mathbf{Z}$ is finite non-empty with doubling constant at most K, then A can be covered by at most $g(K)$ translates of a convex progression P of cardinality at most $f(K)|A|$ and rank at most $d(K)$, for some functions f, g, d.

- For instance, in 2012 Konyagin (refining previous work of Sanders) showed that one can take $d(K) = \log^{3+o(1)} K$ and $f(K) = g(K) = \exp(\log^{3+o(1)} K)$ in this formulation, whereas in the original formulation it is easy to see that $f(K)$ must grow exponentially in K.
- The *Polynomial Freiman–Ruzsa conjecture* over the integers asserts that (in the above formulation) one can take $d(K) = O(\log K)$ and $f(K) = g(K) = O(K^{O(1)})$.
- This remains open.

- Freiman's theorem was extended to arbitrary abelian groups $G = (G, +)$ by Green and Ruzsa in 2007 (with the notion of a convex progression generalized to that of a convex coset progression).
- The Sanders–Konyagin quantitative version of Freiman's theorem extends to this case.
- We will focus on the case of *m*-torsion groups G for some fixed natural number m. These are abelian groups where $mx = 0$ for all $x \in G$.
- A key example are the standard vector spaces $\mathbf{F}_2^n$ for large n; these are the finite 2-torsion groups, and are of particular interest in theoretical computer science.

Marton's PFR conjecture

Freiman–Ruzsa theorem (Ruzsa, 1999)

If G is an m-torsion group and $A \subset G$ is finite non-empty with doubling constant at most K, then A can be covered by at most $g(m, K)$ cosets of a subgroup H of cardinality at most $f(m, K)|A|$, for some functions f, g.

By subdividing the group H, one can always take $f(m, K) = 1$ (at the cost of increasing $g(m, K)$ to $mf(m, K)g(m, K)$).
Let $g_*(m, K)$ denote the optimal value of $g(m, K)$ with $f(m, K) = 1$.

Polynomial Freiman–Ruzsa (PFR) conjecture (Marton, 1999)

$g_*(m, K) \ll_m K^{O_m(1)}$.

(Technically, Marton conjectured $g_*(m, K) \leq K^{O_m(1)}$), but this version can easily be seen to fail for K very close to 1.)

History of results towards PFR:

- Ruzsa (1999): $g_*(m,K) \leq K^2 m^{K^4+1}$.
- Green, T. (2009): $g_*(2,K) \leq 2K^{O(\sqrt{K})}$; for downsets, $g_*(2,K) \leq 2K^{O(1)}$.
- Schoen (2011): $g_*(m,K) \leq \exp(m \exp(O(\sqrt{\log K})))$.
- Sanders (2010): $g_*(m,K) \leq \exp(m \log^{4+o(1)} K)$.
- Konyagin (2012): $g_*(m,K) \leq \exp(m \log^{3+o(1)} K)$.
- GGMT (2023): $g_*(2,K) \leq 2K^{12}$.
- Liao (2023): $g_*(2,K) \leq 2K^{11}$.
- Lean collaboration (2023): Formalized the preceding two results in `Lean`.
- GGMT (2024): $g_*(m,K) \leq (2K)^{O(m^3)}$.
- Liao (2024): $g_*(2,K) \leq 2K^9$.

Thus Marton's PFR conjecture holds for all m.

PFR for *m*-torsion groups does not directly imply PFR for the integers (or vice versa). However, by combining PFR for 2-torsion groups with a previous argument of Green, Manners, and myself, we have a "weak" version of PFR over torsion-free groups:

Weak PFR over $\mathbf{Z}^d$ (GMT 2023 + GGMT 2023)

If $A \subset \mathbf{Z}^d$ is finite non-empty with doubling constant at most K, then A can be covered by $O(K^{O(1)})$ translates of a subspace of $\mathbf{R}^d$ of dimension $O(\log K)$.

Formalized in `Lean` with K^{18} translates and dimension at most $40 \log_2 K$.
PFR over the integers remains a challenging open problem; only a portion of our arguments extend to this case.

By previous work, there are several further consequences of the PFR conjecture. Here are some sample ones:

Approximate homomorphisms close to actual homomorphisms

If $f : \mathbf{F}_2^n \to \mathbf{F}_2^k$ is such that $\mathbf{P}_{x,y \in \mathbf{F}_2^n}(f(x) + f(y) = f(x+y)) \geq 1/K$, then there exists a linear map $g : \mathbf{F}_2^n \to \mathbf{F}_2^k$ such that $\mathbf{P}_{x \in \mathbf{F}_2^k}(f(x) = g(x)) \gg K^{-O(1)}$.

- This is a routine consequence of PFR and the Balog–Szemerédi–Gowers lemma.
- In fact it is equivalent to the $m = 2$ case of PFR (an observation essentially due to Ruzsa).
- Formalized in `Lean` with $\mathbf{P}_{x \in \mathbf{F}_2^n}(f(x) = g(x)) \geq 2^{-172} K^{-146}$.

Polynomial inverse theorem for Gowers U^3 norm

If $f : \mathbf{F}_2^n \to \mathbf{C}$ is 1-bounded with $\|f\|_{U^3(\mathbf{F}_2^n)} \geq 1/K$, then there exists a quadratic polynomial $Q : \mathbf{F}_2^n \to \mathbf{F}_2$ such that $|\mathbf{E}_{x \in \mathbf{F}_2^n} f(x)(-1)^{Q(x)}| \gg K^{-O(1)}$.

- This follows from PFR and arguments of Samorodnitsky (2007).
- Was previously known to be equivalent to the $m = 2$ case of PFR (Lovett 2012; Green–T. 2010).
- Analogous results hold in odd characteristic.
- For the experts: the polynomial Bogulybov conjecture remains open. However, that conjecture is not needed to establish the polynomial U^3 inverse theorem.

Another consequence of PFR is

Sum-product theorem in $\mathbf{R}$ (Mugdal, 2023)

Let $A \subset \mathbf{R}$ be finite non-empty. Then $|mA| + |A^m| \gg |A|^{f(m)}$ for some $f(m)$ that goes to infinity as $m \to \infty$.

A famous conjecture of Erdős and Szemerédi (1983) conjectures that one can take $f(m) = m - \varepsilon$ for any $\varepsilon > 0$.

Methods of proof

- Ruzsa's original arguments were purely combinatorial (or "physical space") in nature, using tools from what we now call Ruzsa calculus, such as the Plünnecke–Ruzsa inequalities and the Ruzsa covering lemma.
- Later works primarily relied on Fourier-analytic methods, as well as versions of the Croot-Sisask lemma. (An exception is the result for downsets, which instead used the method of compressions.)
- Surprisingly, our arguments use no Fourier methods whatsoever, relying instead on entropy methods (in particular, *Shannon entropy inequalities*).

- While the proof crucially requires entropy methods, it is possible to describe the *heuristic* ideas of the proof without reference to entropy.
- A convenient concept in Ruzsa calculus is the Ruzsa distance

$$d[A;B] := \log \frac{|A-B|}{|A|^{1/2}|B|^{1/2}}$$

between two finite non-empty sets A, B.
- This distance is symmetric, non-negative, and satisfies the Ruzsa triangle inequality $d[A;C] \leq d[A;B] + d[B;C]$. (But we caution that $d[A;A] \neq 0$ in general.)
- This distance measures how "commensurable" A and B are.

- For simplicity we work in $\mathbf{F}_2^n$.
- By Ruzsa calculus, PFR is equivalent to the assertion that every K-doubling subset A of lies within $O(\log K)$ (in Ruzsa distance) of a subgroup of $\mathbf{F}_2^n$.
- By an induction on K, the Ruzsa triangle inequality, and previous results on PFR, it would suffice to show that every K-doubling subset A of lies within $O(\log K)$ of a set of doubling constant at most $O(K^{0.99})$ (say).
- Thanks to Ruzsa calculus, many "natural" operations on A will only move the set by $O(\log K)$ in Ruzsa distance.
- So the task is to somehow modify the given K-doubling set A by "natural operations" to improve the doubling constant.

First key example

- Suppose that A is a random subset of a large finite subgroup H of $\mathbf{F}_2^n$, of density $1/K$.
- Then the doubling constant of A is K with high probability.
- However, if we replace A with $A + A$, then we will very likely have replaced A with H, which has doubling constant 1.
- So replacing A by $A + A$ is one of the "natural operations" we would like to perform.

Second key example

- Now suppose that A is the union of K random cosets of a finite subgroup H (of large index).
- Then the doubling constant of A is $\asymp K$ with high probability.
- In this case, replacing A by $A + A$ will likely make the doubling constant worse ($\asymp K^2$ rather than $\asymp K$).
- However, replacing A by $A \cap (A + h)$ for "typical" $h \in A - A$ will usually replace A with a coset of H, bringing the doubling constant down to 1 again.
- So replacing A by $A \cap (A + h)$ is another "natural operation" we would like to perform.

Hybrid example

- Now let A be a random subset of K_1 random cosets of H, of density $1/K_2$, where the cardinality and index of H are both large compared to K_1, K_2.
- Here the doubling constant of A is typically $\asymp K_1 K_2$.
- Replacing A with $A + A$ typically changes the doubling constant to $\asymp K_1^2$.
- Replacing A with $A \cap (A + h)$ typically changes the doubling constant to $\asymp K_2^2$.
- Note that the original doubling constant behaves like the geometric mean of the doubling constant of the two modifications of A.
- Hence, at least one of these operations will improve, or at least not worsen, the doubling constant.

Heuristic argument

- In general, given a finite non-empty set $A \subset G$ and a homomorphism $\pi : G \to H$, the doubling constant of A is *heuristically* at least as large as the doubling constant of $\pi(A)$, times the doubling constant of typical fibers $\pi^{-1}(\{h\})$, $h \in \pi(A)$. Let us informally refer to this as the "fibring inequality".
- The fibring inequality is justified when the fibers $\pi^{-1}(\{h\})$, $h \in \pi(A)$ all have comparable size.
- Near-equality in the fibring inequality is only expected when the fiber sumsets $\pi^{-1}(\{h\}) + \pi^{-1}(\{k\})$ depend "primarily" on $h + k$ rather than on h and k separately.
- Applying this heuristic to $A \times A \subset G^2$ and the addition homomorphism $\pi : (x, y) \mapsto x + y$, we expect that in general, the doubling constant of A is at least the geometric mean of the doubling constant of $A + A$ and of the typical fiber $A \cap (A + h)$.

- This leads to at least one natural operation improving the doubling constant, unless the fibring inequality is close to equality.
- Heuristically, this implies that the sumset of $A \cap (A+h)$ and $A \cap (A+k)$ depend primarily on $h+k$, rather than on h and k separately.
- Alternatively: if $a_1, a_2, a_3, a_4 \in A$, $h = a_2 + a_1$, and $k = a_4 + a_3$, and we fix the value of $h + k = a_2 + a_1 + a_4 + a_3$, then $h = a_2 + a_1$ has no significant influence on the sum $a_1 + a_3$.

- We thus have to handle the “endgame” situation in which, after fixing $a_2 + a_1 + a_4 + a_3$, $a_2 + a_1$ and $a_1 + a_3$ behave like independent random variables.
- **Key observation in characteristic two**: $(a_2 + a_1) + (a_1 + a_3) = a_2 + a_3$ has the same distribution as either $a_2 + a_1$ or $a_1 + a_3$, even after fixing $a_2 + a_1 + a_4 + a_3$.
- Thus, the region where the random variable $a_2 + a_1$ (or $a_1 + a_3$) is concentrated should have quite a small doubling constant.
- In the $m = 2$ case, this provides the final “natural operation” needed to obtain the desired improvement in the doubling constant!

Making things rigorous

- To make this argument rigorous, we should work with pairs A, B of sets rather than a single set A (because we will often need to sum one fiber against another).
- This is a minor technicality that can be dealt with primarily by appropriate notational changes.
- The biggest problem is that the fibring inequality is false in general, due to the variable sizes of fibers $\pi^{-1}(\{h\})$.
- In fact, one can even construct (moderately pathological) examples where a projection $\pi(A)$ has strictly larger doubling constant than A!

- To resolve this problem, we replace sets A with random variables X. The analogue of the logarithm $\log|A|$ of cardinality $|A|$ is then the Shannon entropy

$$\mathbf{H}[X] := \sum_x \mathbf{P}[X = x] \log \frac{1}{\mathbf{P}[X = x]}.$$

- Instead of taking fibers, one works with conditional entropies

$$\mathbf{H}[X|Y] := \sum_y \mathbf{P}[Y = y] \mathbf{H}[X|Y = y].$$

- Heuristically, the entropic formulation makes the "microstate" fibers "essentially" the same size (the Shannon-McMillan-Breiman equipartition theorem).

- Another key notion from information theory is the conditional mutual information

$$\mathbf{I}[X : Y|Z] := \mathbf{H}[X|Z] + \mathbf{H}[Y|Z] - \mathbf{H}[X, Y|Z].$$

- We have the important submodularity inequality

$$\mathbf{I}[X : Y|Z] \geq 0$$

  with equality if and only if X, Y are conditionally independent over Z.
- Thus, conditional mutual information is a quantitative measure of conditional independence.

- The analogue of the logarithm $\log K$ of the doubling constant is the entropic doubling constant

$$\sigma[X] := \mathbf{H}[X + X'] - \mathbf{H}[X]$$

where X' is an independent copy of X.

- Similarly we have the entropic Ruzsa distance

$$d[X; Y] := \mathbf{H}[X' - Y'] - \frac{1}{2}\mathbf{H}[X] - \frac{1}{2}\mathbf{H}[Y]$$

where X', Y' are independent copies of X, Y.

- Many “Ruzsa calculus” inequalities in additive combinatorics have entropic analogues, which can be proven by judicious applications of submodularity.
- For instance, the submodularity inequality

$$\mathbf{I}[X - Y : Z | X - Z] \geq 0$$

can be rearranged (with additional basic entropy facts) to conclude the entropic Ruzsa triangle inequality

$$d[X; Z] \leq d[X; Y] + d[Y; Z].$$

- Similarly, if X_1, X_2 are independent copies of X in G and $\pi : G \to H$ is a homomorphism, the submodularity inequality

$$\mathbf{I}[X_1 + X_2 : \pi(X_1), \pi(X_2) | \pi(X_1 + X_2)] \geq 0$$

gives (among other things) the contraction property

$$\sigma[\pi(X)] \leq \sigma[X]$$

that failed in the combinatorial setting.

- With these tools, one can obtain a rigorous entropic version of the fibring inequality, and make the previous PFR argument rigorous for $m = 2$.
- Many technical optimizations can then be performed to get explicit bounds such as $g_*(2, K) \leq 2K^{12}$ or $g_*(2, K) \leq 2K^{11}$.
- For $m > 2$, one uses a similar strategy, but with (entropic) doubling constant replaced by a "multidistance" relating m different variables $X_1, \dots, X_m$:

$$D[X_1, \dots, X_m] := \mathbf{H}[X_1' + \dots + X_m'] - \frac{1}{m} \sum_{i=1}^{m} \mathbf{H}[X_i],$$

where $X_1', \dots, X_m'$ are independent copies of $X_1, \dots, X_m$ respectively.

- One then creates an $m \times m$ array $X_{i,j}$ of such variables, and shows that it is possible to improve the multidistance by natural operations unless the random variables

$$\sum_{i=1}^{m}\sum_{j=1}^{m} iX_{i,j}, \sum_{i=1}^{m}\sum_{j=1}^{m} jX_{i,j}$$

are almost independent conditioning on $\sum_{i=1}^{m}\sum_{j=1}^{m} X_{i,j}$.
- The key is then to use the m-torsion to note that the difference

$$\sum_{i=1}^{m}\sum_{j=1}^{m} iX_{i,j} - \sum_{i=1}^{m}\sum_{j=1}^{m} jX_{i,j} = \sum_{i=1}^{m}\sum_{j=1}^{m} (i-j)X_{i,j}$$

has the same distribution as either of the two double sums, even after conditioning.

PFR over the integers?

- One can set up the same basic strategy of trying to improve something like the entropic doubling constant through natural operations.
- The problem now is that there is a new example of random variable whose entropic doubling does not improve through such operations: discrete gaussians (concentrated over a large convex progression).
- What is missing is a way to “detect” discrete gaussian structure by purely entropic means (without already assuming PFR).

Lean formalization

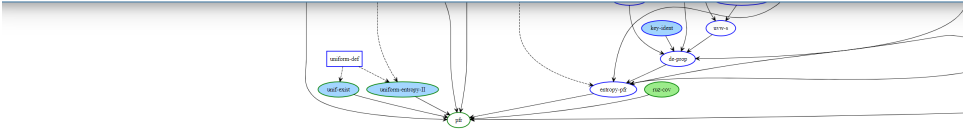
- Shortly after the $m = 2$ case of PFR was established, Yaël Dillies and I launched a project to formalize the proof in the formal proof assistant language `Lean`.
- With many contributions from approximately twenty volunteers, this formalization was completed in three weeks.
- A major component of this formalization was the development of the basic theory of Shannon entropy, which is now in the process of being uploaded to Lean's central math library `Mathlib`.

- The first step in formalization was to create a blueprint.
- This is a human-readable version of the proof (written in a version of LaTeX) that breaks down the proof into many lemmas, linked together by a dependency graph.

Theorem 7.2 (PFR)

If $A \subset \mathbf{F}_2^n$ and $|A+A| \leq K|A|$, then A can be covered by most $2K^{12}$ translates of a subspace H of $\mathbf{F}_2^n$ with $|H| \leq |A|$.

LaTeX *Lean*



- Each node of the graph comes with a human-readable proof of the statement associated to that node, assuming all the results of the dependent nodes.
- Individuals then volunteer to formalize in the proof of selected nodes.
- This can be done in any order and is a highly parallelizable process.

Theorem 7.2 (PFR)

> If $A \subset \mathbf{F}_2^n$ and $|A+A| \leq K|A|$, then A can be covered by most $2K^{12}$ translates of a subspace H of $\mathbf{F}_2^n$ with $|H| \leq |A|$.

Proof ▸

Let U_A be the uniform distribution on A (which exists by Lemma 2.5), thus $H[U_A] = \log|A|$ by Lemma 2.7. By Lemma 2.3 and the fact that $U_A + U_A$ is supported on $A+A$, $H[U_A + U_A] \leq \log|A+A|$. By Definition 3.7, the doubling condition $|A+A| \leq K|A|$ therefore gives

$$d[U_A; U_A] \leq \log K.$$

By Theorem 6.16, we may thus find a subspace H of $\mathbb{F}_2^n$ such that

- One does not need to understand the entire project in order to formalize a single node.
- For instance, much of the work on formalizing the theory of Shannon entropy was done by probabilists with no prior experience in additive combinatorics.

```
theorem PFR_conjecture                                                    source
      {G : Type u_1} [AddCommGroup G] [ElementaryAddCommGroup G 2]
      [Fintype G] [DecidableEq G] {A : Set G} {K : ℝ}
      (h₀A : Set.Nonempty A)
      (hA : ↑(Nat.card ↑(A + A)) ≤ K * ↑(Nat.card ↑A)) :
  ∃ H c,
   ↑(Nat.card ↑c) ≤ 2 * K ^ 12 ∧
    Nat.card ↥H ≤ Nat.card ↑A ∧ A ⊆ c + ↑H
```

- Because `Lean` verifies the validity of all contributed proofs, no prior trust amongst contributors was required.
- This allows for far larger collaborations than traditional math projects.

```
/--   $$ d[X;Y] \geq 0.$$  -/
lemma rdist_nonneg : 0 ≤ d[ X ; μ # Y ; μ' ] := by
  suffices : 0 ≤ 2 * d[ X ; μ # Y ; μ' ]
  . linarith
  have h : |H[X ; μ] - H[Y ; μ']| ≤ 2 * d[X ; μ # Y ; μ' ] := by
    exact diff_ent_le_rdist
  have h' : 0 ≤ |H[X ; μ] - H[Y ; μ']| := by
    exact abs_nonneg (H[X; μ] - H[Y; μ'])
  exact ge_trans h h'
```

- AI tools such as Github Copilot were modestly helpful in the formalization process, essentially serving as an advanced "autocomplete" feature.
- In the future, I expect AI tools to automate more of the tedious steps of proof formalization. Eventually, it may become faster to write a correct formal proof than a correct informal one!

```
/--   $$ d[X;Y] \geq 0.$$   -/
lemma rdist_nonneg : 0 ≤ d[ X ; μ # Y ; μ' ] := by
  suffices : 0 ≤ 2 * d[ X ; μ # Y ; μ' ]
  . linarith
  have h : |H[X ; μ] - H[Y ; μ']| ≤ 2 * d[X ; μ # Y ; μ' ] := by
    rw [abs_of_nonneg (entropy_nonneg _), abs_of_nonneg (entropy_nonneg _)]
    exact diff_ent_le_rdist
  sorry
```

Thanks for listening!