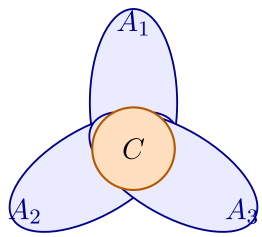
A classroom proof of the improved sunflower lemma

The ALWZ breakthrough in the Rao–Tao spread-family formulation



common core C; disjoint petals $A_i \setminus C$

This note is meant for a one- or two-lecture presentation. It proves the modern bound

$$\mathrm{Sun}(w, r) \leq (Cr \log w)^w$$

for a universal constant C, suppressing constant optimization. The proof uses the spread-family route developed after Alweiss–Lovett–Wu–Zhang: extract a core, prove the residual family is spread, and use random bins to find disjoint petals.

All logarithms are base 2 unless stated otherwise.

Contents

1 Statement and strategy 2

2 Spread families 2

3 Extracting the core 3

4 The spread-family engine 4

5 Proof of the improved bound 4

6 Why the classical proof loses a factor w 5

7 Entropy proof of the single-petal lemma 5

7.1 The refinement inequality 6

7.2 Using refinement to locate one petal 7

8 A five-minute board proof 7

9 Exercises 8

10 References 8

1 Statement and strategy

Definition 1.1 (Sunflower)**.** A family $A_1, \ldots, A_r$ of distinct sets is an *r-sunflower* if there is a set C such that

$$A_i \cap A_j = C \qquad \text{for all } i \neq j.$$

The set C is the *core*. The sets $A_i \setminus C$ are the *petals*. Thus a sunflower is the same thing as a common core together with pairwise disjoint petals.

Let $\mathrm{Sun}(w, r)$ be the least N such that every family of more than N distinct sets, each of size at most w, contains an r-sunflower. Erdős and Rado proved the classical bound

$$\mathrm{Sun}(w, r) \leq w!(r-1)^w.$$

The modern improvement is as follows.

Theorem 1.2 (Improved sunflower lemma)**.** *There is a universal constant C such that for all $w \geq 2$ and $r \geq 2$,*

$$\mathrm{Sun}(w, r) \leq (Cr \log w)^w.$$

The proof has three steps.

1. **Core extraction.** If a small set C is contained in many members of $\mathcal{F}$, put it into the core. Choose C maximally by a weighted popularity score.
2. **Spread petals.** After removing this core, no test set S is too popular. This condition is called *spreadness*.
3. **Random bins.** A spread family is hit by a random set of density $1/(2r)$ with probability at least $1/2$. Partition the universe into $2r$ random bins; in expectation at least r bins contain a petal. Those petals are disjoint.

The entire proof can be summarized as

$$\boxed{\text{large family} \;\Rightarrow\; \text{core} \;+\; \text{spread residual family} \;\Rightarrow\; \text{disjoint petals.}}$$

2 Spread families

Definition 2.1 (R-spread)**.** Let A be a random subset of a finite universe X. It is *R-spread* if for every $S \subseteq X$,

$$\mathbb{P}[S \subseteq A] \leq R^{-|S|}.$$

A finite family $\mathcal{F}$ is R-spread if a uniformly random member $A \in \mathcal{F}$ is R-spread.

Equivalently,

$$|\{A \in \mathcal{F} : S \subseteq A\}| \leq R^{-|S|}|\mathcal{F}| \qquad \text{for every } S \subseteq X.$$

So no fixed t-set lies in more than an R^{-t} fraction of the family.

Remark 2.2. The classical proof mostly controls degrees of single elements. Spreadness controls every small set S at once: elements, pairs, triples, and so on. This all-dimensional control is the reason the base improves from roughly rw to roughly $r \log w$.

Lemma 2.3 (Subsets preserve spreadness)**.** *If A is an R-spread random set and $B \subseteq A$ always, then B is also R-spread.*

Proof. For every S,

$$\mathbb{P}[S \subseteq B] \leq \mathbb{P}[S \subseteq A] \leq R^{-|S|}.$$

□

Example 2.4 (Block-product family)**.** Let $X = B_1 \sqcup \cdots \sqcup B_w$, where each block B_i has size R. Let

$$\mathcal{F} = \{A \subseteq X : |A \cap B_i| = 1 \text{ for every } i \in [w]\}.$$

Then $|\mathcal{F}| = R^w$, and $\mathcal{F}$ is R-spread. Indeed, a fixed set S is contained in a random $A \in \mathcal{F}$ with probability 0 if it uses two elements from one block, and otherwise with probability $R^{-|S|}$.

If $V \sim \text{Bin}(X, p)$, then V contains some member of $\mathcal{F}$ exactly when V hits every block. Hence

$$\mathbb{P}[\exists A \in \mathcal{F} : A \subseteq V] = \left(1 - (1-p)^R\right)^w.$$

For $p \asymp 1/r$, this becomes a positive constant only around $R \asymp r \log w$. Thus the logarithm is not an artifact of the random-set method.

3 Extracting the core

For a family $\mathcal{F}$ and a set C, write

$$\mathcal{F}(C) = \{A \in \mathcal{F} : C \subseteq A\}, \qquad \mathcal{F}_C = \{A \setminus C : A \in \mathcal{F}(C)\}.$$

Think of C as a proposed core and $\mathcal{F}_C$ as the residual petal family.

Lemma 3.1 (Core extraction)**.** *Let $\mathcal{F}$ be a w-uniform family and let $R > 1$. There is a set C such that $\mathcal{F}_C$ is R-spread. If also $|\mathcal{F}| > R^w$, then $|C| < w$ and*

$$|\mathcal{F}_C| > R^{w-|C|}.$$

Proof. Define the weighted popularity

$$Q(C) = R^{|C|}|\mathcal{F}(C)|.$$

Choose C maximizing $Q(C)$.

Let $S \subseteq X$. If $S \cap C \neq \varnothing$, then no member of $\mathcal{F}_C$ contains S, since the elements of C have been removed. Hence the spread inequality is trivial. Assume now that $S \cap C = \varnothing$. Then

$$\begin{aligned} \mathbb{P}_{G \in \mathcal{F}_C}[S \subseteq G] &= \frac{|\mathcal{F}(C \cup S)|}{|\mathcal{F}(C)|} \\ &\leq \frac{R^{|C|}}{R^{|C \cup S|}} = R^{-|S|}, \end{aligned}$$

where the inequality is $Q(C \cup S) \leq Q(C)$.

Thus $\mathcal{F}_C$ is R-spread. Moreover,

$$R^{|C|}|\mathcal{F}(C)| = Q(C) \geq Q(\varnothing) = |\mathcal{F}| > R^w,$$

so $|\mathcal{F}_C| = |\mathcal{F}(C)| > R^{w-|C|}$. If $|C| = w$, then a w-uniform family has at most one member containing C, contradicting $|\mathcal{F}_C| > 1$. Therefore $|C| < w$. □

4 The spread-family engine

The only technical input is the next proposition.

Proposition 4.1 (Single-petal lemma)**.** *There is a universal constant C_0 such that the following holds. Let $\mathcal{G}$ be an R-spread family of sets of size at most m, where $m \geq 2$. Let $V \sim \mathrm{Bin}(X, 1/(2r))$. If*

$$R \geq C_0 r \log m,$$

then

$$\mathbb{P}_V[\exists G \in \mathcal{G} \text{ such that } G \subseteq V] \geq \frac{1}{2}.$$

For the first classroom pass, this is the black box. Section 7 gives the proof idea. The key qualitative point is the parameter $R = O(r \log m)$.

Lemma 4.2 (Spread families contain disjoint members)**.** *There is a universal constant C_1 such that if $\mathcal{G}$ is R-spread, all members of $\mathcal{G}$ have size at most $m \leq w$, and*

$$R \geq C_1 r \log w,$$

then $\mathcal{G}$ contains r pairwise disjoint sets.

Proof. Randomly partition the universe into $2r$ bins

$$X = V_1 \sqcup \cdots \sqcup V_{2r},$$

placing each element independently and uniformly into one bin. Marginally, each V_i has distribution $\mathrm{Bin}(X, 1/(2r))$.

Let E_i be the event that V_i contains some member of $\mathcal{G}$. By the single-petal lemma, $\mathbb{P}[E_i] \geq 1/2$. Hence the expected number of successful bins is

$$\mathbb{E} \sum_{i=1}^{2r} 1_{E_i} = \sum_{i=1}^{2r} \mathbb{P}[E_i] \geq r.$$

Therefore some partition has at least r successful bins. Choose one set from $\mathcal{G}$ inside each of r successful bins. These chosen sets are pairwise disjoint because the bins are disjoint. □

5 Proof of the improved bound

We first prove the theorem for w-uniform families.

Uniform case. Let

$$R = C_1 r \log w,$$

where C_1 is large enough for the previous lemma. Suppose $\mathcal{F}$ is w-uniform and

$$|\mathcal{F}| > R^w.$$

By core extraction, there is a set C such that $\mathcal{F}_C$ is R-spread and $|C| < w$. The members of $\mathcal{F}_C$ have size

$$m = w - |C| \leq w.$$

The spread-family lemma gives r pairwise disjoint members

$$G_1, \ldots, G_r \in \mathcal{F}_C.$$

For each i, choose $A_i \in \mathcal{F}$ with $A_i = C \cup G_i$. Since the G_i are pairwise disjoint,

$$A_i \cap A_j = C \qquad (i \neq j).$$

Thus $A_1, \ldots, A_r$ form an r-sunflower. □

For families of sets of size at most w, pad each set A with fresh dummy elements that appear in no other padded set until it has size exactly w. A sunflower in the padded family projects to a sunflower in the original family, because dummy elements are never shared by two distinct sets.

This proves

$$\mathrm{Sun}(w, r) \leq (Cr \log w)^w$$

for a universal constant C.

6 Why the classical proof loses a factor w

The classical proof says: if there are not r disjoint sets, take a maximal disjoint subfamily. It has at most $r-1$ sets, so its union has size at most $w(r-1)$. Every set in $\mathcal{F}$ intersects this union. By pigeonhole, some element is contained in at least a $1/(w(r-1))$ fraction of the family. Remove that element and induct.

This uses only element degrees. It therefore naturally produces a base of order rw.

The improved proof replaces "some element is popular" by "some test set S is popular". If such an S exists, we put it into the core. If no such S exists, the family is spread, and random sets of density $1/r$ find petals. The loss is then $\log w$, not w.

argument	information controlled	base
Erdős–Rado	single elements	$O(rw)$
ALWZ/Rao–Tao	all test sets S	$O(r \log w)$

7 Entropy proof of the single-petal lemma

This section is optional. It explains the mechanism behind the black-box proposition. Constants are deliberately loose.

Let

$$h(\delta) = \delta \log \frac{1}{\delta} + (1-\delta) \log \frac{1}{1-\delta}$$

be the binary entropy function.

7.1 The refinement inequality

Lemma 7.1 (Refinement inequality)**.** *Let A be an R-spread random subset of X. Let $W \sim \mathrm{Bin}(X, \delta)$ be independent of A. Let (A', W') be a conditionally independent copy of (A, W) subject to*

$$A' \cup W' = A \cup W.$$

Then A' has the same distribution as A, $A' \subseteq A \cup W$, and, if $R\delta > 1$,

$$\mathbb{E}|A' \setminus W| \leq \frac{1 + h(\delta)}{\log(R\delta)} \mathbb{E}|A|.$$

A conditionally independent copy means: reveal $A \cup W$; then sample (A', W') independently from the conditional distribution of (A, W) given this same union.

Entropy proof sketch. Because $A' \cup W' = A \cup W$, we have $A' \subseteq A \cup W$, and therefore

$$A' \setminus W \subseteq A \cap A'.$$

It is enough to bound $\mathbb{E}|A \cap A'|$.

Estimate $H(A, W, A', W')$ in two ways. Conditional independence given the union gives

$$\begin{aligned} H(A, W, A', W') &= H(A, W) + H(A', W' \mid A, W) \\ &= 2H(A, W) - H(A \cup W). \end{aligned}$$

Since A and W are independent, and revealing $A \cup W$ rather than W costs at most $\log(1/\delta)$ bits per element of A,

$$H(A, W, A', W') \geq 2H(A) + H(W) - \log(1/\delta)\mathbb{E}|A|.$$

For the upper bound, reveal the variables in the order

$$A, \quad A \cap A', \quad A', \quad W, \quad W'.$$

The needed estimates are

$$\begin{aligned} H(A \cap A' \mid A) &\leq \mathbb{E}|A|, \\ H(A' \mid A \cap A') &\leq H(A') - (\log R)\mathbb{E}|A \cap A'|, \\ H(W \mid A, A', A \cap A') &\leq H(W) - \log(1/\delta)\mathbb{E}|A' \setminus A|, \\ H(W' \mid A, A', W, A \cap A') &\leq h(\delta)\mathbb{E}|A'|. \end{aligned}$$

The second line is the spreadness input. Learning a subset $S \subseteq A'$ saves at least $|S| \log R$ bits on average. The third line uses $A' \setminus A \subseteq W$. The last line is the entropy of the random deleted part $A' \setminus W'$ once A' is known.

Adding the four estimates and using $A' \overset{d}{=} A$ gives

$$\begin{aligned} H(A, W, A', W') \leq 2H(A) &+ H(W) - \log(1/\delta)\mathbb{E}|A| \\ &+ (1 + h(\delta))\mathbb{E}|A| - \log(R\delta)\mathbb{E}|A \cap A'|. \end{aligned}$$

Comparing with the lower bound yields

$$\log(R\delta)\mathbb{E}|A \cap A'| \leq (1 + h(\delta))\mathbb{E}|A|.$$

Since $A' \setminus W \subseteq A \cap A'$, the lemma follows. □

7.2 Using refinement to locate one petal

Proof idea for the single-petal lemma. Let A_0 be a uniformly random member of the R-spread family $\mathcal{G}$. Then $\mathbb{E}|A_0| \le m$.

Choose

$$L = \lceil \log(2m) \rceil, \qquad \delta = \frac{1}{4rL}.$$

Let $W_1, \dots, W_L$ be independent samples from $\mathrm{Bin}(X, \delta)$, and set $U_i = W_1 \cup \cdots \cup W_i$. The union U_L has density at most $L\delta = 1/(4r)$, so it can be coupled as a subset of $V \sim \mathrm{Bin}(X, 1/(2r))$.

By applying the refinement inequality successively, one constructs random sets $A_1, \dots, A_L$, each distributed as A_0, with

$$A_i \subseteq A_0 \cup U_i \qquad \text{and} \qquad \mathbb{E}|A_i \setminus U_i| \le \gamma^i \mathbb{E}|A_0|,$$

where

$$\gamma = \frac{1 + h(\delta)}{\log(R\delta)}.$$

At step i, apply the refinement inequality to the remaining fragment $A_i \setminus U_i$, conditionally on the already exposed U_i. This fragment is still R-spread because it is a subset of an R-spread random set.

If the absolute constant in $R \ge C_0 r \log m$ is large enough, then $R\delta \ge 16$. Since $h(\delta) \le 1$,

$$\gamma \le \frac{2}{\log 16} = \frac{1}{2}.$$

Therefore

$$\mathbb{E}|A_L \setminus U_L| \le 2^{-L} m \le \frac{1}{2}.$$

Because $|A_L \setminus U_L|$ is integer-valued,

$$\mathbb{P}[A_L \subseteq U_L] \ge \frac{1}{2}.$$

Since $U_L \subseteq V$ under the coupling and A_L is distributed as a random member of $\mathcal{G}$, the random set V contains a member of $\mathcal{G}$ with probability at least $1/2$. □

8 A five-minute board proof

For a compressed classroom ending, write the following.

1. Define R-spread: $\Pr[S \subseteq A] \le R^{-|S|}$ for every test set S.
2. Given $|\mathcal{F}| > R^w$, choose C maximizing

   $$Q(C) = R^{|C|} |\{A \in \mathcal{F} : C \subseteq A\}|.$$

   Then $\mathcal{F}_C$ is R-spread.
3. Spread lemma: if $R \ge C_0 r \log w$, then a random set of density $1/(2r)$ contains a member of $\mathcal{F}_C$ with probability at least $1/2$.

4. Randomly partition the universe into $2r$ bins. The expected number of bins containing a residual petal is at least r.

5. For some partition, r bins succeed. Pick one residual petal in each successful bin. They are disjoint. Add C back to get an r-sunflower.

9 Exercises

Exercise 9.1. Show that $A_1, \ldots, A_r$ form a sunflower if and only if the sets $A_i \setminus C$ are pairwise disjoint for $C = \bigcap_i A_i$.

Exercise 9.2. Prove the padding reduction from families of sets of size at most w to w-uniform families.

Exercise 9.3. In the core-extraction lemma, explain why $Q(C \cup S) \leq Q(C)$ is exactly the inequality needed to prove R-spreadness of $\mathcal{F}_C$.

Exercise 9.4. Prove the random-partition lemma without using independence between the events E_i.

Exercise 9.5. For the block-product family, prove that

$$\mathbb{P}[\exists A \in \mathcal{F} : A \subseteq V] = \left(1 - (1-p)^R\right)^w.$$

Use this to show that the random-set strategy cannot generally work with $R \ll r \log w$ when $p \asymp 1/r$.

Exercise 9.6. Assume only that every element appears in at most a $1/R$ fraction of a w-uniform family $\mathcal{F}$. Show that a greedy argument finds r disjoint sets if $R > w(r-1)$. Compare this with the spread-family threshold.

10 References

The original breakthrough is due to Alweiss, Lovett, Wu, and Zhang. Rao gave a shorter coding-theoretic proof; Tao gave a Shannon-entropy proof; Bell, Chueluecha, and Warnke recorded a small adjustment giving the clean $O(r \log w)^w$ form. These notes follow the spread-family presentation and leave constants unoptimized.

References

[1] P. Erdős and R. Rado. Intersection theorems for systems of sets. *Journal of the London Mathematical Society* 35 (1960), 85–90.

[2] R. Alweiss, S. Lovett, K. Wu, and J. Zhang. Improved bounds for the sunflower lemma. *Annals of Mathematics* 194(3) (2021), 795–815. https://annals.math.princeton.edu/2021/194-3/p05

[3] A. Rao. Coding for sunflowers. *Discrete Analysis* 2020:2. https://discreteanalysisjournal.com/article/11887-coding-for-sunflowers

[4] T. Tao. The sunflower lemma via Shannon entropy. Blog post, 2020. https://terrytao.wordpress.com/2020/07/20/the-sunflower-lemma-via-shannon-entropy/

[5] T. Bell, S. Chueluecha, and L. Warnke. Note on sunflowers. *Discrete Mathematics* 344(7) (2021), article 112367. https://doi.org/10.1016/j.disc.2021.112367

[6] M. Stoeckl. Lecture notes on recent improvements for the sunflower lemma. 2022. https://mstoeckl.com/notes/research/sunflower_notes.html