

# Coding for Sunflowers

Anup Rao  
University of Washington  
anuprao@cs.washington.edu

January 8, 2020

## Abstract

A sunflower is a family of sets that have the same pairwise intersections. We simplify a recent result of Alweiss, Lovett, Wu and Zhang that gives an upper bound on the size of every family of sets of size  $k$  that does not contain a sunflower. We show how to use the converse of Shannon’s noiseless coding theorem to give a cleaner proof of a similar bound.

## 1 Introduction

A  $p$ -sunflower is a family of  $p$  sets whose pairwise intersections are identical. How large can a family of sets of size  $k$  be if the family does not contain a  $p$ -sunflower? Erdős and Rado [ER60] were the first to pose and answer this question. They showed that any family with more than  $(p-1)^k \cdot k!$  sets of size  $k$  must contain a  $p$ -sunflower. This fundamental fact has many applications in mathematics and computer science [ES92, Raz85, FMS97, GM07, GMR13, Ros14, RR18, LZ19, LSZ19].

After nearly 60 years, the correct answer to this question is still not known. There is a family of  $(p-1)^k$  sets of size  $k$  that does not contain a  $p$ -sunflower, and Erdős and Rado conjectured that their lemma could be improved to show that this is essentially the extremal example. Recently, Alweiss, Lovett, Wu and Zhang [ALWZ19] made substantial progress towards resolving the conjecture. They showed that  $(\log k)^k \cdot (p \log \log k)^{O(k)}$  sets ensure the presence of a  $p$ -sunflower. Subsequently, Frankton, Kahn, Narayanan and Park [FKNP19] improved the counting methods developed by [ALWZ19] to prove a conjecture of Talagrand [Tal10] regarding monotone set systems.

In this work, we give simpler proofs for these results. Our proofs rely on an encoding argument inspired by a similar encoding argument used in [ALWZ19, FKNP19]. The main novelty is our use of Shannon’s noiseless coding theorem [Sha48, Kra49] to reason about the efficiency of the encoding, which turns out to avoid complications that show up when using vanilla counting. We show<sup>1</sup>:

**Theorem 1.** *There is a universal constant  $\alpha > 1$  such that every family of more than  $(\alpha p \log(pk))^k$  sets of size  $k$  must contain a  $p$ -sunflower.*

Let  $r(p, k)$  denote the quantity  $\alpha p \log(pk)$ . We say<sup>2</sup> that a sequence<sup>3</sup> of sets  $S_1, \dots, S_\ell \subset [n]$  of size  $k$  is  $r$ -spread if for every non-empty set  $Z \subset [n]$ , the number of elements of the sequence that

<sup>1</sup>In this paper, all logarithms are computed base 2.

<sup>2</sup>A similar concept was first used by Talagrand [Tal10].

<sup>3</sup>Here we state the results for sequences of sets because some applications require the ability to reason about sequences that may repeat sets.

contain  $Z$  is at most  $r^{k-|Z|}$ . We prove that for an appropriate choice of  $\alpha$ , the following lemma holds:

**Lemma 2.** *If a sequence of more than  $r(p, k)^k$  sets of size  $k$  is  $r(p, k)$ -spread, then the sequence must contain  $p$  disjoint sets.*

As far as we know, it is possible that Lemma 2 holds even when  $r(p, k) = O(p)$ . Such a strengthening of Lemma 2 would imply the sunflower conjecture of Erdős and Rado. Lemma 2 easily implies Theorem 1: we proceed by induction on  $k$ . When  $k = 1$ , the theorem holds, since the family contains  $p$  distinct sets of size 1. For  $k > 1$ , if the sets are not  $r$ -spread, then there is a non-empty set  $Z$  such that more than  $r^{k-|Z|}$  of the sets contain  $Z$ . By induction, and since  $r(p, k)$  can only increase with  $k$ , the family of sets contains a  $p$ -sunflower. Otherwise, if the sets are  $r(p, k)$ -spread, Lemma 2 guarantees the presence of a  $p$ -sunflower.

It only remains to prove Lemma 2. In fact, we prove something much stronger: a small random set is very likely to contain *some* set of an  $r$ -spread family of sets.

## 2 Random sets and $r$ -spread families

To prove Lemma 2, we need to understand the extent to which a small random set  $W \subseteq [n]$  contains some set of a large family of sets of size  $k$ . To that end, it is convenient to use the following definition:

**Definition 3.** *Given  $S_1, \dots, S_\ell \subseteq [n]$ , for  $x \in [\ell]$  and  $W \subseteq [n]$ , let  $\chi(x, W)$  be equal to  $S_y \setminus W$ , where  $y \in [\ell]$  is chosen to minimize  $|S_y \setminus W|$  among all choices with  $S_y \subseteq S_x \cup W$ . If there are multiple choices for  $y$  that minimize  $|S_y \setminus W|$ , let  $y$  be the smallest one.*

Observe that the definition makes sense even if  $S_1, \dots, S_\ell$  are not all distinct. When  $U \subseteq W$ , we have  $|\chi(x, U)| \geq |\chi(x, W)|$ . Moreover,  $\chi(x, W) = \emptyset$  if and only if there is an index  $y$  for which  $S_y \subseteq W$ . Our main technical lemma shows that if a long sequence of sets is  $r$ -spread, then  $|\chi(X, W)|$  is likely to be small for a random  $X$  and a random small set  $W$ :

**Lemma 4.** *There is a universal constant  $\beta > 1$  such that the following holds. Let  $0 < \gamma, \epsilon < 1/2$ . If  $r = r(k, \gamma, \epsilon) = \beta \cdot (1/\gamma) \cdot \log(k/\epsilon)$ , and  $S_1, \dots, S_\ell \subseteq [n]$  is an  $r$ -spread sequence of at least  $r^k$  sets of size  $k$ ,  $X \in [\ell]$  is uniformly random, and  $W \subseteq [n]$  is a uniformly random set of size at least  $\gamma n$  independent of  $X$ , then  $\mathbb{E}[|\chi(X, W)|] < \epsilon$ . In particular,  $\Pr_W[\exists y, S_y \subseteq W] > 1 - \epsilon$ .*

This lemma is of independent interest — it is relevant to several applications in theoretical computer science [Ros14, LSZ19]. Before we prove Lemma 4, let us see how to use it to prove Lemma 2.

*Proof of Lemma 2.* Set  $\gamma = 1/(2p)$ ,  $\epsilon = 1/p$ . Then  $r = r(k, \gamma, \epsilon) = r(p, k)$ . Let  $W_1, \dots, W_p$  be a uniformly random partition of  $[n]$  into sets of size at least  $\lfloor n/p \rfloor$ . So, each set  $W_i$  is of size at least  $\lfloor n/p \rfloor \geq \gamma n$ . By symmetry and linearity of expectation, we can apply Lemma 4 to conclude that

$$\mathbb{E}_{X, W_1, \dots, W_p} [|\chi(X, W_1)| + \dots + |\chi(X, W_p)|] = \mathbb{E}_{X, W_1} [|\chi(X, W_1)|] + \dots + \mathbb{E}_{X, W_p} [|\chi(X, W_p)|] < \epsilon p = 1.$$

Since  $|\chi(X, W_1)| + \dots + |\chi(X, W_p)|$  is a non-negative integer, there must be some fixed partition  $W_1, \dots, W_p$  for which

$$\mathbb{E}_X [|\chi(X, W_1)| + \dots + |\chi(X, W_p)|] = 0.$$

This can happen only if the family contains  $p$  disjoint sets.  $\square$

Next, we briefly describe a technical tool from information theory, before turning to prove Lemma 4.

### 3 Prefix-free encodings

A *prefix-free* encoding is a map  $E : [t] \rightarrow \{0, 1\}^*$  into the set of all binary strings, such that if  $i \neq j$ ,  $E(i)$  is not a prefix of  $E(j)$ . Another way to view such an encoding is as a map from the set  $[t]$  to the vertices of the infinite binary tree. The encoding is prefix-free if  $E(i)$  is never an ancestor of  $E(j)$  in the tree.

Shannon [Sha48] proved that one can always find a prefix-free encoding such that the expected length of the encoding of a random variable  $X \in [t]$  exceeds the entropy of  $X$  by at most 1. Conversely, every encoding must have average length that is at least as large as the entropy. For our purposes, we only need the converse under the uniform distribution. The proof is short, so we include it here. All logarithms are taken base 2.

**Lemma 5.** *Let  $E : [t] \rightarrow \{0, 1\}^*$  be any prefix-free encoding, and  $\ell_i$  be the length of  $E(i)$ . Then  $(1/t) \cdot \sum_{i=1}^t \ell_i \geq \log t$ .*

*Proof.* We have

$$\log t - (1/t) \cdot \sum_{i=1}^t \ell_i = (1/t) \cdot \sum_{i=1}^t \log(t \cdot 2^{-\ell_i}) \leq \log \left( \sum_{i=1}^t 2^{-\ell_i} \right),$$

where the inequality follows from the concavity of the logarithm function. The fact that this last quantity is at most 0 is known as Kraft's inequality [Kra49]. Consider picking a uniformly random binary string longer than all the encodings. Because the encodings are prefix-free, the probability that this random string contains the encoding of some element of  $[t]$  as a prefix is exactly  $\sum_{i=1}^t 2^{-\ell_i}$ . So, this number is at most 1, and the above expression is at most 0.  $\square$

### 4 Proof of Lemma 4

We shall prove that there is a constant  $\kappa$  such that the following holds. For each integer  $m$  with  $0 \leq m \leq r\gamma/\kappa$ , if  $W$  is a uniformly random set of size at least  $\kappa mn/r$ , then  $\mathbb{E} [|\chi(X, W)|] \leq k(12/13)^m$ . By the choice of  $r(k, \gamma, \epsilon)$ , setting  $m = \lfloor r\gamma/\kappa \rfloor$ , we get that when  $W$  is a set of size at least  $\gamma n$ ,  $\mathbb{E} [|\chi(X, W)|] \leq k(12/13)^{\alpha \log(k/\epsilon)/\kappa} < \epsilon$  for  $\alpha$  chosen large enough.

We prove the bound by induction on  $m$ . When  $m = 0$ , the bound holds trivially. When  $m > 0$ , sample  $W = U \cup V$ , where  $U, V$  are uniformly random disjoint sets,  $|U| = u = \lceil \kappa(m-1)n/r \rceil$ , and  $|V| = v \geq \kappa n/r - 1 \geq \kappa nr/2$ . Note that we always have  $\kappa/2 \leq (rv/n) \leq 2\kappa$ .

It is enough to prove that for all fixed choices of  $U$ ,

$$13 \cdot \mathbb{E}_{V, X} [|\chi(X, W)|] \leq 12 \cdot \mathbb{E}_X [|\chi(X, U)|].$$

So, fix  $U$ . If  $\chi(x, U)$  is empty for any  $x$ , then we have  $\mathbb{E}_{V, X} [|\chi(X, W)|] = \mathbb{E}_X [|\chi(X, U)|] = 0$ , so there is nothing to prove. Otherwise, we must have  $\mathbb{E}_X [|\chi(X, U)|] \geq 1$ , since  $|\chi(x, U)| \geq 1$  for all  $x$ . The number of possible pairs  $(V, X)$  is at least  $r^k \cdot \binom{n-u}{v}$ . Our bound will follow from using Lemma 5. We give a prefix-free encoding of  $(V, X)$  as follows:

1. For each  $A \subseteq \chi(X, U)$ , with  $|A| = |\chi(X, W)|$ , define

$$\tau(A, X, V) = \{y \in [\ell] : A \subseteq \chi(y, U) \subseteq W \cup \chi(X, U), |\chi(y, U)| = |\chi(X, U)|\},$$

and define

$$\phi(X, V) = r^k \cdot (32v/n)^{|\chi(X, U)|} \cdot (vr/n)^{-|\chi(X, W)|}.$$

The first case is that for all  $A$ ,  $|\tau(A, X, V)| \leq \phi(X, V)$ . In this case, the first bit of the encoding is set to 0, and we proceed to encode  $(V, X)$  like this:

- (a) Encode  $|\chi(X, U)|$ . It suffices to use a trivial encoding of this integer: we encode it with the string  $0^{|\chi(X, U)|}1$ , which has length  $|\chi(X, U)| + 1$ .
- (b) Encode  $W \cup \chi(X, U)$ . Since  $U$  has been fixed, there are

$$\binom{n-u}{v} + \dots + \binom{n-u}{v+|\chi(X, U)|} \leq \binom{n-u+|\chi(X, U)|}{v+|\chi(X, U)|} \leq \binom{n-u}{v} \cdot (n/v)^{|\chi(X, U)|}$$

choices for this set. So, the encoding has length at most

$$\log \left( \binom{n-u}{v} \cdot (n/v)^{|\chi(X, U)|} \right) + 1.$$

- (c) Let  $j$  be such that  $\chi(j, U) \subseteq W \cup \chi(X, U)$ , and  $|\chi(j, U)|$  is minimized. If there are multiple choices for  $j$  that achieve the minimum, let  $j$  be the smallest one.  $X$  is a potential candidate for  $j$ , so we must have  $|\chi(j, U)| \leq |\chi(X, U)|$ . Encode  $\chi(X, U) \cap \chi(j, U)$ . Since  $j$  is determined, this takes at most  $|\chi(X, U)|$  bits.
- (d) We have already encoded  $\chi(j, U) \cap \chi(X, U) \subseteq S_X$ . We claim that this set must have size at least  $|\chi(X, W)|$ . Indeed,  $\chi(j, U) = S_h \setminus U$  for some set  $S_h$  of the  $r$ -spread sequence. We have

$$S_h \setminus U = \chi(j, U) \subseteq \chi(X, U) \cup W,$$

so

$$S_h \subseteq \chi(X, U) \cup W \subseteq S_X \cup W.$$

By the definition of  $\chi(X, W)$ , this implies that

$$|\chi(X, W)| \leq |S_h \setminus W| = |S_h \setminus W \cap \chi(X, U) \setminus W| \leq |\chi(j, U) \cap \chi(X, U)|,$$

as claimed. Now, since  $|\tau(A, X, V)| \leq \phi(X, V)$  for all  $A$  of size  $|\chi(X, W)|$ , we can encode  $X$  using a binary string of length at most

$$\log \left( r^k \cdot (32v/n)^{|\chi(X, U)|} \cdot (vr/n)^{-|\chi(X, W)|} \right) + 1.$$

- (e) Because  $X$  has been encoded,  $\chi(X, U)$  is also determined. Encode  $W \cap \chi(X, U)$ . Together with  $W \cup \chi(X, U)$ , this determines  $W$ , and so  $V$ . This last step takes  $|\chi(X, U)|$  bits.

Combining all of the above steps, and using the fact that  $|\chi(X, U)| \geq 1$  and  $vr/n \geq \kappa/2$ , the total length of the encoding in this case is at most

$$\begin{aligned} & \log \left( r^k \cdot \binom{n-u}{v} \right) + 8 \cdot |\chi(X, U)| - \log(vr/n) \cdot |\chi(X, W)| + 4 \\ & \leq \log \left( r^k \cdot \binom{n-u}{v} \right) + 12 \cdot |\chi(X, U)| - 13 \cdot |\chi(X, W)|, \end{aligned}$$

for  $\kappa$  chosen large enough.

2. In the second case there is a set  $A \subseteq \chi(X, U)$  of size  $|\chi(X, W)|$  such that  $|\tau(A, X, V)| > \phi$ . Then the first bit of the encoding is set to 1, and we proceed like this:

- (a) Encode  $X$ . This takes at most  $\log r^k + 1$  bits.
- (b) Now  $\chi(X, U)$  is determined. Encode the set  $A$  promised above. This takes at most  $|\chi(X, U)|$  bits.
- (c) We claim that at this point, the number of candidates for  $V$  is at most  $\binom{n-u}{v} \cdot 16^{-|\chi(X, U)|}$ . Indeed, consider the following random experiment. Choose a set  $B$  uniformly at random from the collection of sets satisfying  $A \subseteq B \subseteq \chi(X, U)$ , and then sample  $V \subseteq [n] \setminus U$  uniformly at random. Consider the collection of  $y \in \tau(A, X, V)$  for which  $B = \chi(y, U) \cap \chi(X, U)$ . Define  $N(A, B, X, V) = |\{y \in [\ell] : B = \chi(y, U) \cap \chi(X, U)\} \cap \tau(A, X, V)|$ . We have

$$\mathbb{E}_{B, V} [N(A, B, X, V)] \leq \mathbb{E}_B \left[ r^{k-|B|} \cdot \left( \frac{v}{n-u-k} \right)^{|\chi(X, U)|-|B|} \right].$$

This is because there are at most  $r^{k-|B|}$  sets of size  $|\chi(X, U)|$  containing  $B$ , and for each one, the probability that it is included in  $V$  is at most  $(v/(n-u-k))^{|\chi(X, U)|-|B|}$ . By the choice of  $u$ , we have  $n-u-k \geq n/3$ . So, we continue to bound:

$$\leq \mathbb{E}_B \left[ r^{k-|B|} \cdot \left( \frac{3v}{n} \right)^{|\chi(X, U)|-|B|} \right] \leq r^k \cdot (3v/n)^{|\chi(X, U)|} \cdot (vr/n)^{-|\chi(X, W)|}.$$

The last inequality holds because  $|B| \geq |\chi(X, W)|$ . On the other hand, we have

$$\Pr[|\tau(A, X, V)| > \phi(X, V)] \cdot 2^{-|\chi(X, U)|} \cdot \phi(X, V) \leq \mathbb{E}_{B, V} [N(A, B, X, V)],$$

since  $B$  takes each value with probability at least  $2^{-|\chi(X, U)|}$ . By the definition of  $\phi$ ,

$$\Pr[|\tau(A, X, V)| > \phi(X, V)] \leq 16^{-|\chi(X, U)|}.$$

So, we can encode  $V$  at a cost of

$$\log \binom{n-u}{v} - 4 \cdot |\chi(X, U)| + 1.$$

Overall, the cost of carrying out the encoding in the second case is at most (by adding  $13 \cdot (|\chi(X, U)| - |\chi(X, W)|)$ ):

$$\log \left( r^k \binom{n-u}{v} \right) - 3 \cdot |\chi(X, U)| + 2 \leq \log \left( r^k \binom{n-u}{v} \right) + 12|\chi(X, U)| - 13|\chi(X, W)|.$$

Note that this last bound is exactly the same as the bound we obtained in the case that  $\tau(A, X, V) \leq \phi(X, V)$ .

Now, it only remains to apply Lemma 5. The expected length of the encoding cannot be less than  $\log \left( r^k \binom{n-u}{v} \right)$ . This implies that  $13 \mathbb{E}[|\chi(X, W)|] \leq 12 \mathbb{E}[|\chi(X, U)|]$ , as required.

## 5 Acknowledgements

Thanks to Ryan Alweiss, Shachar Lovett, Kewen Wu and Jiapeng Zhang for many useful comments. Thanks to Sivaramakrishnan Natarajan Ramamoorthy, Siddharth Iyer and Paul Beame for useful conversations. Thanks to the editor and reviewers of Discrete Analysis for insightful comments.

## References

- [ALWZ19] R. Alweiss, S. Lovett, K. Wu, and J. Zhang. Improved bounds for the sunflower lemma. *arXiv:1908.08483*, 2019.
- [ER60] P. Erdős and R. Rado. Intersection theorems for systems of sets. *Journal of London Mathematical Society*, 35:85–90, 1960.
- [ES92] P. Erdős and A. Sárközy. Arithmetic progressions in subset sums. *Discrete Mathematics*, 102(3):294–264, 1992.
- [FKNP19] K. Frankston, J. Kahn, B. Narayanan, and J. Park. Thresholds versus fractional expectation-thresholds. *arXiv:1910.13433*, 2019.
- [FMS97] G. S. Frandsen, P. B. Miltersen, and S. Skyum. Dynamic word problems. *J. ACM*, 44(2):257–271, 1997.
- [GM07] A. Gál and P. B. Miltersen. The cell probe complexity of succinct data structures. *Theor. Comput. Sci*, 379(3):405–417, 2007.
- [GMR13] P. Gopalan, R. Meka, and O. Reingold. Dnf sparsification and a faster deterministic counting algorithm. *Computational Complexity*, 22(2):275–310, 2013.
- [Kra49] L. G. Kraft. A device for quantizing, grouping, and coding amplitude-modulated pulses. Master’s thesis, Massachusetts Institute of Technology, 1949.
- [LSZ19] S. Lovett, N. Solomon, and J. Zhang. From dnf compression to sunflower theorems via regularity. In A. Shpilka, editor, *34th Computational Complexity Conference, CCC 2019, July 18-20, 2019, New Brunswick, NJ, USA*, volume 137 of *LIPICs*, pages 5:1–5:14. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2019.
- [LZ19] S. Lovett and J. Zhang. Dnf sparsification beyond sunflowers. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019*, pages 454–460, New York, NY, USA, 2019. ACM.
- [Raz85] A. A. Razborov. Lower bounds on the monotone complexity of some Boolean functions. *Doklady Akademii Nauk SSSR*, 281(4):798–801, 1985.
- [Ros14] B. Rossman. The monotone complexity of  $k$ -clique on random graphs. *SIAM Journal on Computing*, 43(1):256–279, 2014.
- [RR18] S. N. Ramamoorthy and A. Rao. Lower bounds on non-adaptive data structures maintaining sets of numbers, from sunflowers. In *33rd Computational Complexity Conference*, volume 102 of *LIPICs*, pages 27:1–27:16, 2018.

- [Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948. Monograph B-1598.
- [Tal10] M. Talagrand. Are many small sets explicitly small? In L. J. Schulman, editor, *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 13–36. ACM, 2010.