Electronic Colloquium on Computational Complexity, Report No. 82 (2018)

Sunflowers and Quasi-sunflowers from Randomness Extractors

Xin Li[*]
Johns Hopkins University, Baltimore

Shachar Lovett[†]
University of California, San Diego

Jiapeng Zhang[‡]
University of California, San Diego

Abstract

The Erdős-Rado sunflower theorem (Journal of Lond. Math. Soc. 1960) is a fundamental result in combinatorics, and the corresponding sunflower conjecture is a central open problem. Motivated by applications in complexity theory, Rossman (FOCS 2010) extended the result to quasi-sunflowers, where similar conjectures emerge about the optimal parameters for which it holds.

In this work, we exhibit a surprising connection between the existence of sunflowers and quasi-sunflowers in large enough set systems, and the problem of constructing certain randomness extractors. This allows us to re-derive the known results in a systemic manner, and to reduce the relevant conjectures to the problem of obtaining improved constructions of the randomness extractors.

1 Introduction

Let $\mathcal{F}$ be a collection of sets from some universe X. A common theme and extensively studied phenomenon in combinatorics is the following: if the cardinality of $\mathcal{F}$ (when $\mathcal{F}$ is finite) or the density of $\mathcal{F}$ (when $\mathcal{F}$ is infinite) is large enough, then some nice patterns will occur in $\mathcal{F}$. Well known examples of this kind include (1) Szemerédi's theorem [Sze75], which asserts that all subsets of the natural numbers of positive density contain arbitrarily long arithmetic progressions; (2) Ramsey's theorem [ES35], which asserts that if one colors the edges of a large enough complete graph with a finite number of colors, then there must exist a monochromatic clique of a certain size; and (3) the Erdős-Rado sunflower theorem [ER60], which asserts that a large enough collection of subsets with bounded size of a universe must contain a large sunflower.[1]

The study of these problems has resulted in many important tools (e.g., Szemerédi's regularity lemma [Sze78] and the probabilistic method), which have found wide applications not only in combinatorics, but also in computer science. Conversely, ideas from computer science have also influenced related research in combinatorics quite often. For example, the first two problems we mentioned above, Szemerédi's theorem and Ramsey's theorem, are intimately connected to the area of pseudorandomness in theoretical computer science. Indeed, by constructing a certain sparse pseudorandom subset of natural numbers and proving an appropriate Szemerédi-type theorem with respect to that subset, a celebrated result of Green and Tao [GT08] shows that prime numbers

[*] Email: lixints@cs.jhu.edu. Research supported by NSF award CCF-1617713.

[†] Email: slovett@cs.ucsd.edu. Research supported by NSF CCF-1614023.

[‡] Email: jpeng.zhang@gmail.com. Research supported by NSF CCF-1614023.

[1] A sunflower is a collection of sets whose pairwise intersection is constant, which we will formally define shortly.

ISSN 1433-8092

contain arbitrarily long arithmetic progressions. As for Ramsey's theorem, a recent line of work on randomness extractors [Li15], [CZ16], [BADTS17], [Coh17], [Li17] give highly explicit constructions of Ramsey graphs that almost match the probabilistic bound [Erd47].

In this paper we study the sunflower theorem and its related variants. We show that again there is an intimate connection to randomness extractors. In fact, using the techniques from randomness extractors, we build a general proof framework that can unify the sunflower theorem and its variant known as the quasi-sunflower lemma [Ros10]. Furthermore, any improvement in the analysis of the extractors will lead to improvements in the lemmas. We now begin our formal discussion of the sunflower theorem and the quasi-sunflower lemma.

Sunflowers. An r-*sunflower* is defined to be a collection of r sets from some universe X, such that the intersection of any two sets is the same (which can be the empty set). Choose any collection $\mathcal{F}$ of sets from X, the main question of interest is how large $\mathcal{F}$ needs to be in order to ensure that there is a r-sunflower in $\mathcal{F}$. Erdős and Rado proved the following theorem.

Theorem 1.1 ([ER60]). *Let $\mathcal{F}$ be an arbitrary family of sets from some universe X, where each set in $\mathcal{F}$ has size w. If $|\mathcal{F}| > w!(r-1)^w$ then $\mathcal{F}$ contains an r-sunflower.*

They also conjectured that the bound on $|\mathcal{F}|$ can be replaced by c_r^w where c_r is a constant that only depends on r for every $r > 0$. This conjecture is one of the most well known open problems in combinatorics, which remains open today despite a lot of research.

The sunflower theorem has applications in computer science, such as proving strong lower bounds for monotone circuits [Raz85]. In addition, a paper by Alon, Shpilka and Umans [ASU12] relates the sunflower conjecture and its variants to possible approaches of achieving fast matrix multiplication. Recently, following breakthrough results that prove the strong Cap Set Conjecture [CFLPP17], [SEG17], a weaker version of the sunflower conjecture known as the Erdős-Szemeredi Sunflower Conjecture is also proved. However, the general conjecture still remains open.

Quasi-sunflowers. Motivated by the applications of sunflowers in proving monotone circuit lower bounds, quasi-sunflowers are introduced by Rossman [Ros10] to prove monotone circuit lower bounds for the k-clique problem on random graphs. We denote by $\mathcal{P}(X)$ the family of all subsets of a finite set X.

Definition 1.2 ([Ros10]). Let X be a finite set, $\mathcal{S} \subseteq \mathcal{P}(X)$ be a family of sets such that $|\mathcal{S}| \geq 2$. Denote $Y = \bigcap_{T \in \mathcal{S}} T$. For $p, \gamma \in [0,1]$, $\mathcal{S}$ is said to be a (p,γ)-quasi-sunflower if for a random set $W \subseteq X$, with each element of X present in W independently with probability p,

$$\Pr\left[\exists T \in \mathcal{S}, (T \setminus Y) \subseteq W\right] \geq 1 - \gamma.$$

In the same paper, Rossman also proved the following quasi-sunflower lemma, which says there is always a quasi-sunflower in a large family of subsets.

Lemma 1.3 (Quasi-sunflower lemma, [Ros10]). *Let $\mathcal{F}$ be a family of sets over a universe X each of size w. If $|\mathcal{F}| \geq w! \cdot (1.71 \log(1/\gamma)/p)^w$, then $\mathcal{F}$ contains a (p,γ)-quasi-sunflower.*

Besides the original application, Rossman's quasi-sunflower lemma was also used by Gopalan, Meka and Reingold [GMR13] to study the problem of DNF sparsification. Given a DNF formula f on n variables, there are two natural ways to measure the complexity of f: the number of clauses (also called size) $s(f)$, and the maximum width of a clause $w(f)$. It is easy to show that any DNF of small size can be approximated well by another DNF of small width, by truncating clauses of larger width. Gopalan et al. [GMR13] used Rossman's quasi-sunflower lemma to show the reverse direction, that any DNF with small width can also be approximated well by another DNF with small size. In particular, they showed that any width w DNF formula can be ε-approximated by another DNF formula with size at most $(w \log(1/\varepsilon))^{O(w)}$. This kind of sparsification has applications in constructing pseudorandom generators and approximate counting the number of satisfying assignments for DNF formulas.

Similar to the sunflower conjecture, one can also ask whether the bound on $\mathcal{F}$ in the quasi-sunflower lemma can be improved. For example, if one can improve the bound to $(O(\log(1/\gamma)/p))^w$ then it is also possible to improve the ε-approximation of DNF formula in [GMR13] to have size $(\log(1/\varepsilon))^{O(w)}$.

1.1 Our contribution

We provide a general framework to prove both the sunflower theorem and the quasi-sunflower lemma. In fact, we reduce both of these problems to the construction of a certain type of randomness extractors. To state our results, we first formally define the notions that are going to be used in our extractors.

Definition 1.4. Let D be a distribution over a sample space X. The min-entropy of D is defined as

$$\mathcal{H}_\infty(D) = \min_x \left\{ \log \left(\frac{1}{\Pr[D = x]} \right) \right\}.$$

Definition 1.5 (Block min-entropy source). A distribution $X = (X_1, \cdots, X_m)$ where each $X_i \in \{0,1\}^n$ is an (m, n, k) *block min-entropy* source if for every non-empty subset $S \subseteq [m]$, the joint distribution of $(X_i : i \in S)$ has min-entropy at least $k|S|$.

We note that the definition of block min-entropy sources was initiated in [GLM+16] as a tool to prove lifting theorems in communication complexity.

Definition 1.6 (Block min-entropy extractor). A function $E : \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^d$ is a (k, ε, d, s) *block min-entropy* extractor if for any $m, n \in \mathsf{N}$ and any (m, n, k) block min-entropy source $X = (X_1, \cdots, X_m)$, we have that

$$(E(X_1, R_1), \cdots, E(X_m, R_m)) \approx_\varepsilon U_{dm}.$$

Here, each $R_i \in \{0,1\}^s$ is an independent uniform random string, U_{dm} is the uniform distribution on dm bits, and $\approx_\varepsilon$ means ε close in the statistical distance. If in addition we have that

$$(E(X_1, R_1), R_1, \cdots, E(X_m, R_m), R_m) \approx_\varepsilon U_{(d+s)m},$$

then we say that the function E is a *strong* (k, ε, d, s) block min-entropy extractor.

We also define a weaker object called a disperser.

Definition 1.7 (Block min-entropy disperser). A function

$$E : \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^d$$

is a (k, ε, d, s) *block min-entropy disperser* if for any $m, n \in \mathsf{N}$ and any (m, n, k) block min-entropy source $X = (X_1, \cdots, X_m)$, we have that

$$|\mathsf{Supp}(E(X_1, R_1), \cdots, E(X_m, R_m))| \geq (1 - \varepsilon)2^{dm}.$$

Here, each $R_i \in \{0,1\}^s$ is an independent uniform random string, and Supp means the support of the distribution. If in addition there exists at least one fixing of $R_1 = r_1, \cdots, R_m = r_m$ such that

$$|\mathsf{Supp}(E(X_1, r_1), \cdots, E(X_m, r_m))| \geq (1-\varepsilon)2^{dm},$$

then we say that the function E is a *strong* (k, ε, d, s) block min-entropy disperser.

In this paper, we make connections between the block min-entropy disperser and (quasi-)sunflower structures. Formally, we prove the following theorem.

Theorem 1.8. *Suppose that there exists a strong $(k, 0, d, s)$ block min-entropy disperser, $E : \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^d$ for any (w, n, k) block min-entropy source. Then the following holds.*

Let $\mathcal{F}$ be a family of sets where each set has size w. Assume that $|\mathcal{F}| \geq 2^{(k+2)w}$. Then:

(i) $\mathcal{F}$ contains a 2^d-sunflower.

(ii) $\mathcal{F}$ contains a $\left(p, w(1-p)^{2^d}\right)$-quasi-sunflower.

Observe that the seed length s of the extractor does not play a part in the conclusion of Theorem 1.8. We then show that we can construct strong block min-entropy extractors and strong zero-error block min-entropy dispersers. Specifically, we have the following theorem.

Theorem 1.9. *There is a constant $c > 1$ such that for any $m, n, k \in \mathsf{N}$ with $k \geq c \log m$, we have:*

- *There is an explicit strong (k, ε, d, s) block min-entropy extractor $E : \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^d$ for (m, n, k) block min-entropy sources, where $s = n$, $d = k/c$ and $\varepsilon = 2^{-\Omega(k)}$.*
- *There is an explicit strong $(k, 0, d, s)$ block min-entropy disperser $E : \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^d$ for (m, n, k) block min-entropy sources, where $s = n$, $d = k/c$.*

Combined with Theorem 1.8, this gives the sunflower theorem and the quasi-sunflower lemma.

Corollary 1.10 (Sunflower theorem, this paper). *There is a constant $c > 1$ such that for any family of sets $\mathcal{F}$ each of size w and any $r > 1$, if $|\mathcal{F}| \geq (wr)^{cw}$, then $\mathcal{F}$ contains a r-sunflower.*

Corollary 1.11 (Quasi-sunflower lemma, this paper). *There is a constant $c > 1$ such that for any family of sets $\mathcal{F}$ each of size w, if $|\mathcal{F}| \geq 2^{2w} \cdot \left(\frac{w + \log(1/\gamma)}{p}\right)^{cw}$, then $\mathcal{F}$ contains a (p, γ)-quasi-sunflower.*

1.2 Overview of the techniques

Our reduction from sunflower/quasi-sunflower problems to block min-entropy dispersers is as follows. Suppose the family $\mathcal{F} \subseteq \mathcal{P}(X)$ for some set X, where each set in $\mathcal{F}$ has size w. We first show that without loss of generality we can assume $\mathcal{F}$ has a *normal form.*

Definition 1.12 (Normal form). Let X be a finite set and let $\mathcal{F} = \{U_i\}_{i \in I}$ be a family of subsets of X. We say that $\mathcal{F}$ is w-normal if

- For each $U \in \mathcal{F}$, the size of U is w.
- There is a disjoint partition $X_1, \cdots, X_w$ of X such that for every $U \in \mathcal{F}$, we have $|X_j \cap U| = 1$ for each $j \in [w]$.

Consider the uniform distribution over $\mathcal{F}$ of a normal form. There are two possible cases:

- **Case 1:** there is a subset S which appears in many sets of $\mathcal{F}$, that is
$$\big|\{U \in \mathcal{F} : S \subseteq U\}\big| \geq \big|\mathcal{F}\big|/\kappa^{|S|},$$
where κ is a parameter to be determined.
- **Case 2:** every set S does not appear in too many sets of $\mathcal{F}$.

In case 1, S is already like a core in a sunflower or quasi-sunflower, thus we can apply induction on the sub-family $\mathcal{F}_S := \{U \setminus S : (U \in \mathcal{F}) \wedge (S \subseteq U)\}$. In case 2, the condition basically implies that the distribution is relatively flat, which equivalently translates into a block min-entropy source as we defined above. One can naturally imagine that the worst case situation here is that the distribution is actually the uniform distribution over $X_1 \times \cdots \times X_w$, and we show that indeed this is the case by using our zero-error block min-entropy disperser. It is then easy to see that in the worst case, the empty set is a quasi-sunflower, or one can choose a sunflower with size 2^d (the support size in the output of the disperser) whose core is the empty set.

1.3 The role of extractors in our reduction

One can view the block min-entropy extractor/disperser used in our reduction as a gadget, which reduces the sunflower/quasi-sunflower problem in the general case to the much easier case of a uniform distribution (or full support) on $X_1 \times \cdots \times X_w$. This is similar to the role of extractors in recent works that showed lifting theorems from query complexity to communication complexity [GLM+16], and linear programming lower bounds for constraint satisfaction problems [KMR17].

In fact, the extractors used in these works are essentially the same as the extractors used in this work (although in this work we need to show that the extractor/disperser is strong, while in [GLM+16] and [KMR17] this is not necessary), and the barriers for further improvement are also similar. Specifically, in all such constructions one needs the min-entropy $k \geq c \log m$ for some constant $c > 1$, where m is equal to the size of the sets (i.e., w) in our applications. It is unknown if this dependence on m is necessary for a block min-entropy extractor/disperser to exist. If one can remove the dependence of k on m (even at the price of decreasing the output of the extractor/disperser), then our reduction will give improved bounds for both the sunflower problem and the quasi-sunflower problem. In particular, by Theorem 1.8 we will be able to show that any family $\mathcal{F}$ of subsets with $|\mathcal{F}| \geq (g(r))^w$ contains a r-sunflower where $g(r)$ is a function on r, and thus prove the sunflower conjecture. It may also lead to a bound of $(O(\log(1/\gamma)/p))^w$ for $\mathcal{F}$ to contain a (p, γ)-quasi-sunflower, and thus improving the DNF sparsification in [GMR13]. Similarly, removing such a dependence will lead to further improvements in lifting theorems and linear programming lower bounds, as shown in [GLM+16] and [KMR17]. In conclusion, we believe that the study of block min-entropy extractors is an important question that needs further investigation.

1.4 Further discussions

Discussions about the sunflower conjecture. In this paper, we show that for any set system $\mathcal{F}$ of size $|\mathcal{F}| \geq w^{cw}$ for some constant $c > 1$, it contains a 3-sunflower. Furthermore, we show that any set system $\mathcal{F}$ with the following Lipschitz condition, must contain three pairwise disjoint sets.

Definition 1.13 (Lipschitz condition). Given a collection of sets $\mathcal{F}$ and $r > 0$. We say it is r-Lipschitz, if for any subset S,
$$\big|\{U \in \mathcal{F} : S \subseteq U\}\big| \geq \big|\mathcal{F}\big|/r^{|S|}.$$

Corollary 1.14. *There is a constant $c > 0$, such that for any w-normal set system $\mathcal{F}$, if $\mathcal{F}$ is w^c-Lipschitz then it contains w pairwise disjoint sets.*

This w^c-Lipschitz condition actually comes from the requirement of our disperser that $k \geq c \log w$. As discussed above, it is interesting to ask whether this is necessary. In particular, there may be a way to improve this corollary without using dispersers. We make the following conjecture, which implies the sunflower conjecture.

Conjecture 1.15 (Disjoint sets conjecture). *For any $r \geq 3$, there exists a constant $c_r > 1$, such that for any set system $\mathcal{F}$, if $\mathcal{F}$ is c_r-Lipschitz then it contains r pairwise disjoint sets.*

As the disjoint sets conjecture seems hard (it implies the sunflower conjecture), we also make the following simpler conjecture, which is of independent interest.

Conjecture 1.16 (2-disjoint sets conjecture). *There exists a constant $c > 1$, such that for any set system $\mathcal{F}$, if $\mathcal{F}$ is c-Lipschitz then it contains 2 pairwise disjoint sets.*

Discussions about quasi-sunflowers. In this paper, we also study quasi-sunflower structures. In particular, we have the following corollary. Below, we use the notation $O_{p,\gamma}(\cdot)$ to hide the specific dependency on the parameters p, γ, which is of less interest to us.

Corollary 1.17. *There is a constant $c > 0$ such that, for any w-normal family $\mathcal{F}$, if $\mathcal{F}$ is r-Lipschitz where $r = (O_{p,\gamma}(w))^c$, then the empty set is a (p, γ)-quasi-sunflower for $\mathcal{F}$.*

It seems the corollary can be further improved. We make the follow conjecture.

Conjecture 1.18. *There is a constant $c > 0$ such that, for any w-normal family $\mathcal{F}$, if $\mathcal{F}$ is r-Lipschitz where $r = (O_{p,\gamma}(\log w))^c$, then the empty set is a (p, γ)-quasi-sunflower for $\mathcal{F}$.*

The reason for the $\log w$ term is the following example, which we believe is the worst instance for quasi-sunflower structures. Fix $p = \gamma = 1/2$ for convenience. Let $X_1, \ldots, X_w$ be w disjoint sets each of size $c \log w$ for some small enough $c > 0$. Define the collection of sets as $\mathcal{F} := X_1 \times \cdots \times X_w$. Then $\mathcal{F}$ does not contain a (p, γ)-quasi-sunflower.

We note that proving our conjectures, or even improving our corollaries will lead to interesting improvements on the sunflower theorem or the quasi-sunflower lemma. This will in turn lead to improvements in other applications such as DNF sparsification, constructing pseudorandom generators for DNF formulas, and approximate counting the number of satisfying assignments for DNF formulas.

2 Preliminaries

We first review some basic definitions in probability.

Definition 2.1. Let D be a distribution over a sample space X. Its entropy is

$$\mathcal{H}(D) = \sum_x \Pr[D = x] \cdot \log \left(\frac{1}{\Pr[D = x]} \right).$$

Its min-entropy is

$$\mathcal{H}_\infty(D) = \min_x \left\{ \log \left(\frac{1}{\Pr[D = x]} \right) \right\}.$$

Its max-entropy is

$$\mathcal{H}_0(D) = \log |\{\mathsf{Supp}(D)\}|.$$

Definition 2.2 (Statistical distance). Let D_0 and D_1 be distributions over a finite sample space X. The statistical distance between D_0 and D_1 is defined as

$$\mathrm{dist}(D_0, D_1) = \frac{1}{2} \sum_{x \in X} \big| \Pr[D_0 = x] - \Pr[D_1 = 1] \big|.$$

3 A construction of block min-entropy extractor

We use the following well-known extractor based on the inner product function [CG88]. We denote by $\mathbb{F}_q$ the finite field on q elements. When $q = 2^\ell$ we identify $\mathbb{F}_q$ with $\{0,1\}^\ell$ and $\mathbb{F}_q^t$ with $\{0,1\}^{t\ell}$.

Theorem 3.1 **([CG88]).** *Let $t, \ell \geq 1$ and take $q = 2^\ell, n = t\ell$. Let X, Y be independent sources on $\mathbb{F}_q^t \cong \{0,1\}^n$ with min-entropy k_1, k_2 respectively. Let* IP *be the inner product function over the field $\mathbb{F}_q$. Then:*

$$\mathrm{dist}\left((\mathrm{IP}(X,Y), X), (U_\ell, X)\right) \leq \varepsilon \quad \textit{and} \quad \mathrm{dist}\left((\mathrm{IP}(X,Y), Y), (U_\ell, Y)\right) \leq \varepsilon$$

where $\varepsilon = 2^{\frac{-(k_1+k_2-n-\ell)}{2}}$.

Now we can construct a block min-entropy extractor as follows. Given parameters n, k, choose a field $\mathbb{F}_q$ such that $q = 2^\ell$ with $\ell = \alpha k$ for some constant $0 < \alpha < 1$ to be determined later. Without loss of generality we assume that $n = \ell t$ for some integer t. We view $X \in \{0,1\}^n$ as a vector in $\mathbb{F}_q^t$ and choose a uniform independent seed $R \in \{0,1\}^n \cong \mathbb{F}_q^t$.

A block min-entropy extractor

1. Given parameters m, n, k let q, t be as described above.
2. Sample $(x_1, \ldots, x_m)$ from the block min-entropy distribution $X = (X_1, \ldots, X_m) \in (\mathbb{F}_q^t)^m$.
3. Uniformly sample $(R_1, \ldots, R_m) \in (\mathbb{F}_q^t)^m$.
4. Output $Z := (\mathrm{IP}(x_1, R_1), \ldots, \mathrm{IP}(x_m, R_m))$.

We are now ready to prove the following theorem.

Theorem 1.9 (restated). *Let $X = (X_1, \cdots, X_m)$ be an (m, n, k) block min-entropy source. Let $Z \in \{0,1\}^{\ell m}$ be the output of the above block min-entropy extractor applied to X. There exists a constant $c > 1$ such that if $k \geq c \log m$, then the following holds foran error $\varepsilon = 2^{-\Omega(k)}$:*

- *With probability $1 - \varepsilon$ over the fixing of the seed $(R_1, \ldots, R_m)$,*

  $$\left| \Pr[Z = z] - 2^{-\ell m} \right| \leq \varepsilon \cdot 2^{-\ell m} \qquad \forall z \in \{0,1\}^{\ell m}.$$

  In particular, in such cases $\mathcal{H}_0(Z) = \ell m$

- $\mathrm{dist}\left((Z, R_1, \cdots, R_m), (U, R_1, \cdots, R_m)\right) \leq 2\varepsilon$.

Proof. Note that we have a joint distribution $(X_1, \cdots, X_m)$ that has block min-entropy k. The output of the local extractor applied to $(X_1, \cdots, X_m)$, using m independent uniform seeds $(R_1, \cdots, R_m)$, is a distribution $(Z_1, \cdots, Z_m)$ over $\{0,1\}^{\ell m} = \mathbb{F}_q^m$ where $Z_i = \mathrm{IP}(X_i, R_i)$ for each i.

For any fixing of the seed $(R_1 = r_1, \cdots, R_m = r_m)$, the distribution $(Z_1, \cdots, Z_m)$ is a deterministic function of $(X_1, \cdots, X_m)$, and we will view this distribution as a function $\mathcal{D} : \{0,1\}^{\ell m} \to [0,1]$ where the image of each input is its associated probability in the distribution. We now write this function in its Fourier basis:

$$\mathcal{D}(z) = \sum_{S \subseteq [\ell m]} \hat{\mathcal{D}}(S) \chi_S(z),$$

where $z = (z_1, \cdots, z_m) \in \{0,1\}^{\ell m}$, $\chi_S(z) = (-1)^{\sum_{i \in S} z(i)} \in \{+1, -1\}$, and

$$\hat{\mathcal{D}}(S) = 2^{-\ell m} \cdot \sum_z \mathcal{D}(z) \chi_S(z) = 2^{-\ell m} \cdot \mathbb{E}_{z \sim \mathcal{D}}[\chi_S(z)].$$

Here we use $z(i)$ to stand for the i'th *bit* of the string z. This is to distinguish between the notation z_i, which referes to the i'th *block* of the string z, that contains ℓ bits.

Note that $\hat{\mathcal{D}}(\emptyset) = 2^{-\ell m}$ since $\mathcal{D}$ is a probability distribution. Thus we have that $\forall z \in \{0,1\}^{\ell m}$,

$$\left| \mathcal{D}(z) - 2^{-\ell m} \right| = \left| \sum_{S \subseteq [\ell m], S \neq \emptyset} \hat{\mathcal{D}}(S) \chi_S(z) \right| \leq \sum_{S \subseteq [\ell m], S \neq \emptyset} \left| \hat{\mathcal{D}}(S) \right|.$$

Note that for any $S \subseteq [\ell m]$, $\chi_S(Z)$ corresponds to the parity of a subset of the bits in Z. For each $Z_j, j \in [m]$, this parity may or may not involve any bits in Z_j. We will be interested in the number of j's such that $\chi_S(Z)$ involves at least one bit from Z_j, and we call this number $\Delta(S)$. Note that $\Delta(\emptyset) = 0$ and $1 \leq \Delta(S) \leq m$ for any $S \neq \emptyset$.

We now have the following lemma.

Lemma 3.2. *If $\Delta(S) = h$, then with probability $1 - 2^{\frac{-h(1-\alpha)k}{4}}$ over the fixing of the seed $(R_1 = r_1, \cdots, R_m = r_m)$, we have that $|\hat{\mathcal{D}}(S)| \leq 2 \cdot 2^{-\ell m} 2^{\frac{-h(1-\alpha)k}{4}}$.*

Proof. Without loss of generality assume that the Z_j's from which $\chi_S(Z)$ involves at least one bit are $(Z_1, \cdots, Z_h)$. Note that for any $Z_i \in \{0,1\}^\ell = \mathbb{F}_q$, any parity of the bits of Z_i corresponds exactly to the first bit of $a \cdot Z_i$ viewed as a vector in $\{0,1\}^\ell$, for some $a \in \mathbb{F}_q$ and the operation $\cdot$ is multiplication in the field $\mathbb{F}_q$. Moreover this correspondence is a bijection in the sense that different parities correspond to different elements $a \in \mathbb{F}_q$. The special case of parity over the empty set corresponds to the case of $a = 0$. Thus, $\sum_{i \in S} Z(i)$ corresponds to the first bit of $\sum_{j \in [h]} a_j Z_j$ viewed as a vector in $\{0,1\}^\ell$, for some non-zero $\{a_j \in \mathbb{F}_q : j \in [h]\}$. Note that

$$\sum_{j \in [h]} a_j Z_j = \sum_{j \in [h]} a_j \mathrm{IP}(X_j, R_j) = \sum_{j \in [h]} \mathrm{IP}(a_j X_j, R_j) = \mathrm{IP}((a_1 X_1, \cdots, a_h X_h), (R_1, \cdots, R_h)).$$

Since each $a_j \neq 0$ the transformation from $(x_1, \cdots, x_h)$ to $(a_1 x_1, \cdots, a_h x_h)$ is a bijection. Thus we know the distribution $(a_1 X_1, \cdots, a_h X_h)$ has min-entropy kh, while $(R_1, \cdots, R_h)$ has min-entropy nh. Thus by Theorem 3.1 applied over the field $\mathbb{F}_{2^{\ell h}}$ we have that

$$\mathrm{dist}\left((\sum_{j \in [h]} a_j Z_j, R_1, \cdots, R_m), (U_{\ell h}, R_1, \cdots, R_m) \right) \leq 2^{\frac{-(h(k-\ell))}{2}} = 2^{\frac{-h(1-\alpha)k}{2}}.$$

In particular, as $\chi_S(Z)$ is the first bit of $\sum_{j\in[h]} a_j Z_j$, we have

$$\mathrm{dist}\left(\chi_S(Z), R_1, \cdots, R_m), (U_1, R_1, \cdots, R_m)\right) \leq 2^{\frac{-h(1-\alpha)k}{2}}.$$

By Markov's inequality this means that with probability $1 - 2^{\frac{-h(1-\alpha)k}{4}}$ over the fixing of the seed $R = (R_1, \ldots, R_m)$, we have $|\hat{\mathcal{D}}(S)| = |2^{-\ell m} \cdot \mathbb{E}_{z\sim\mathcal{D}}[\chi_S(z)]| \leq 2 \cdot 2^{-\ell m} 2^{\frac{-h(1-\alpha)k}{4}}$. □

Next, note that the number of S with $\Delta(S) = h$ is $\binom{m}{h}(2^\ell - 1)^h \leq 2^{(\ell+\log m)h}$. Recall that $\ell = \alpha k$ and $k \geq c \log m$. We can choose the constants α, c such that $2^{\ell+\log m} 2^{\frac{-(1-\alpha)k}{4}} \leq 2^{-\frac{k}{8}}$. Now we have as long as $k \geq 8$,

$$\sum_{h=1}^{m} \binom{m}{h} (2^\ell - 1)^h 2^{\frac{-h(1-\alpha)k}{4}} \leq \sum_{h=1}^{m} 2^{-\frac{hk}{8}} \leq 2^{-\frac{k}{8}+1}.$$

Set $\varepsilon = 2^{-\frac{k}{8}+2} = 2^{-\Omega(k)}$. By the union bound we have that with probability at least $1 - \varepsilon$ over the fixing of the seed $(R_1 = r_1, \cdots, R_m = r_m)$, for every $S \neq \emptyset$ with $\Delta(S) = h$, $|\hat{\mathcal{D}}(S)| \leq 2 \cdot 2^{-\ell m} 2^{\frac{-h(1-\alpha)k}{4}}$. Thus for any such seed we have that

$$\left|\mathcal{D}(z) - 2^{-\ell m}\right| \leq \sum_{S\subseteq[\ell m], S\neq\emptyset} \left|\hat{\mathcal{D}}(S)\right| \leq \varepsilon \cdot 2^{-\ell m}.$$

This concludes the proof of the first part of Theorem 1.9. For the second part, notice that conditioned on the fixing of any seed $R_1, \ldots, R_m$, with probability $1 - \varepsilon$ the statistical distance is at most ε, and otherwise it is trivially bounded by 1. So overall the statistical distance between $(Z, R_1, \cdots, R_m)$ and $(U_{\ell m}, R_1, \cdots, R_m)$ is at most 2ε. □

4 Compressing set systems by the block min-entropy extractor

In this section, we focus on the set systems that satisfy the Lipschitz condition, and show a compression operator for such set systems. Our compression is based on the block min-entropy extractor. We first show that it suffices to consider w-normal set systems (see Definition 1.12).

Lemma 4.1. *Let $\mathcal{F}$ be a family of sets such that each set has size w. Then there exists a w-normal sub-family $\mathcal{F}'$ of $\mathcal{F}$ with $|\mathcal{F}'| \geq |\mathcal{F}|/2^{2w}$.*

Proof. Let $U \in \mathcal{F}$ be a set, and let $X_1, \ldots, X_w$ be a random partition of X. Then

$$\Pr_{X_1,\cdots,X_w}[\forall j \in [w], |U \cap X_j| = 1] = \frac{w!}{w^w}.$$

Then by an average argument, there is a partition $(X_1, \cdots, X_w)$ such that

$$|\{U \in \mathcal{F} : \forall j \in [w], |U \cap X_j| = 1\}| \geq |\mathcal{F}| \cdot \frac{w!}{w^w}$$

The claim then follows since $\frac{w!}{w^w} \geq 2^{-2w}$. □

Now we can focus on normal set systems. Given a finite set X, we denote by X_p the distribution over subsets $W \subset X$, where each $x \in X$ appears in W independently with probability p.

Lemma 4.2. *Let $u \geq w$. Let c be the constant from theorem 1.9. Then for every w-normal set system which is u^c-Lipschitz (recall Definition 1.13), it holds that*

$$\Pr_{W \sim X_p}[\exists U \in \mathcal{F}, U \subseteq W] \geq 1 - w(1-p)^u.$$

To prove this lemma, we first define a "worst case" instance, and then show that all other instances behave better than this case. Let $X_1^*, \cdots, X_w^*$ be w disjoint sets each of size u. Define the family $\mathcal{U}^*$ as

$$\mathcal{U}^* = \left\{\{x_1, \cdots, x_w\} : \forall j \in [w], x_j \in X_j^*\right\}.$$

Claim 4.3. *Let $\mathcal{U}^*$ as defined above. Then*

$$\Pr_{W \sim X_p}[\exists U \in \mathcal{U}^*, U \subseteq W] \geq 1 - w(1-p)^u.$$

Proof. By the definition of $\mathcal{U}^*$, we have that

$$\begin{aligned}\Pr_W[\forall U \in \mathcal{U}^*, U \not\subseteq W] =& \Pr_W[\exists j \in [w], X_j \cap W = \emptyset] \\ \leq& \sum_{j \in [w]} \Pr_W[X_j \cap W = \emptyset] \\ =& w(1-p)^u.\end{aligned}$$

□

Let X, Y be finite sets, $h : X \to Y$ a map. Given a set $U \subset X$ define $h = \{h(x) : x \in U\} \subset Y$. Given a family $\mathcal{F} \subseteq \mathcal{P}(X)$ define $h(\mathcal{F}) \subseteq \mathcal{P}(Y)$ as

$$h(\mathcal{F}) = \{h : U \in \mathcal{F} \text{ and } h \text{ is injective on } U\}.$$

Lemma 4.4. *Let X and Y be sets, $h : X \to Y$ a map, $\mathcal{F} \subset \mathcal{P}(X)$. Then*

$$\Pr_{W_Y \sim Y_p}[\exists U \in h(\mathcal{F}), U \subseteq W_Y] \leq \Pr_{W_X \sim X_p}[\exists U \in \mathcal{F}, U \subseteq W_X].$$

Proof. Without loss of generality, we can assume the map h is surjective, because elements $y \in Y \setminus h(X)$ do not affect the events. If $|Y| = |X|$ then h is a bijection and hence $\mathcal{F}$ and $h(\mathcal{F})$ are the same, up to renaming the elements. So, assume $|Y| < |X|$. It suffices to prove the lemma for the case that $|Y| = |X| - 1$, as the general case follows from applying this case iteratively (namely, decompose h as a sequence of maps, each reduces the domain size by one).

So, assume $|Y| = |X| - 1$. In this case, there is a unique pair $x_1, x_2 \in X$ such that $h(x_1) = h(x_2) = y$. We may assume without loss of generality (by renaming the elements of Y) that h is the identity map on $X' = X \setminus \{x_1, x_2\}$. This allows us to jointly sample (W_X, W_Y) as follows. Sample $W' \sim X'_p, W'_X \sim \{x_1, x_2\}_p, W'_Y \sim \{y\}_p$ and set $W_X = W' \cup W'_X, W_Y = W' \cup W'_Y$. We will show that for every fixed $W' = w'$,

$$\Pr_{W_Y \sim Y_p}[\exists U \in h(\mathcal{F}), U \subseteq W_Y \mid W' = w'] \leq \Pr_{W_X \sim X_p}[\exists U \in \mathcal{F}, U \subseteq W_X \mid W' = w']. \tag{1}$$

The lemma then follows by averaging over W'.

To that end, fix W'. Let $\mathcal{F}' = \{U \setminus X' : U \in \mathcal{F}, (U \cap X') \subset W'\}$. Note that $\mathcal{F}' \subseteq \mathcal{P}(\{x_1, x_2\})$. Similarly, define $\mathcal{F}'' = \{U \setminus X' : U \in \langle(\mathcal{F}), (U \cap X') \subset W'\}$. Note that $\mathcal{F}'' \subseteq \mathcal{P}(\{y\})$. Equation (1) is equivalent to

$$\Pr_{W'_Y \sim \{y\}_p}[\exists U \in \mathcal{F}'', U \subseteq W'_Y] \leq \Pr_{W'_X \sim \{x_1, x_2\}_p}[\exists U \in \mathcal{F}', U \subseteq W'_X]. \tag{2}$$

We verify Equation (2) by a case analysis.

(i) If $\mathcal{F}''$ is empty then the LHS of Equation (2) is 0, while the RHS is non-negative.

(ii) If $\emptyset \in \mathcal{F}''$ then $\emptyset \in \mathcal{F}'$. In this case, both the LHS and RHS of Equation (2) equal 1.

(iii) If $\mathcal{F}'' = \{\{y\}\}$ then either $\{x_1\} \in \mathcal{F}'$ or $\{x_2\} \in \mathcal{F}'$. In either case, the LHS of Equation (2) equals p, while the RHS is at least p. □

We now prove Lemma 4.2. Let $\mathcal{F}$ be a family of sets that satisfies the assumptions. We will show there is a function h such that $h(\mathcal{F}) = \mathcal{U}^*$. The extractor from 1.9, with an appropriate choice of seed, provides such a function h.

Proof of Lemme 4.2. Let $\mathcal{F}$ be a w-normal family of sets that satisfies the Lipchitz condition. We first define the function h. Since $\mathcal{F}$ is a w-normal set, there exists a partition of X to $X_1, \ldots, X_w$ such that for each $U \in \mathcal{F}$ and $j \in [w]$, $|X_j \cap U| = 1$.

Define the sample space as $X_1 \times \cdots \times X_w$. With a slight abuse to use the notations, we identify $\mathcal{F} \subseteq \mathcal{P}(X_1 \times \cdots \times X_w)$, and let D be a uniform distribution over $\mathcal{F}$. Since $\mathcal{F}$ is u^c-Lipschitz, the distribution D is a $(w, \log X, k)$ block min-entropy source with $k = c \log u \geq c \log w$. Then by Theorem 1.9, there exists seeds $r_1, \ldots, r_w$ such that $(\mathrm{IP}(D_1, r_1), \ldots, \mathrm{IP}(D_w, r_w))$ has full support, where $D = (D_1, \ldots, D_w)$. Note that the output of $\mathrm{IP}(\cdot, \cdot)$ is in $\{0,1\}^{k/c} \cong [u]$. We can now define h as follows:

$$h(x) = (\mathrm{IP}(x, r_j), j) \qquad \forall x \in X_j.$$

Note that by definition, h is injective on any $U \in X_1 \times \cdots \times X_w$. We identify elements of $\mathcal{U}^*$ with $\{(a_1, 1), \ldots, (a_w, w)\}$ with $a_i \in [u]$. Thus $h(\mathcal{F}) = \mathcal{U}^*$. The lemma now follows from Lemma 4.4 and Claim 4.3. □

We will also need the following lemma.

Lemma 4.5. *Let $u \geq w$. Let c be the constant from theorem 1.9. Then for every w-normal set system $\mathcal{F}$ which is u^c-Lipschitz (recall Definition 1.13), it holds that $\mathcal{F}$ contains u pairwise disjoint sets.*

Proof. The proof is very similar to the proof of Lemma 4.2. There is a map h for which $h(\mathcal{F}) = \mathcal{U}^*$. Note that $\mathcal{U}^*$ contains u pairwise disjoint sets, $U'_1, \ldots, U'_u$. By definition, $U'_i = h(U_i)$. But then also $U_1, \ldots, U_u$ must be pairwise disjoint. □

4.1 Sunflowers and quasi-sunflowers from compression

Now we can prove Theorem 1.8.

Theorem 1.8 (restated). *Suppose that there exists a strong $(k, 0, d, s)$-block min-entropy disperser, $E : \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^d$ for any (w, n, k)-block min-entropy source. Then the following holds.*

Let $\mathcal{F}$ be a family of sets where each set has size w. Assume that $|\mathcal{F}| \geq 2^{(k+2)w}$. Then:

(i) $\mathcal{F}$ contains a 2^d-sunflower.

(ii) $\mathcal{F}$ contains a $\left(p, w(1-p)^{2^d}\right)$-quasi-sunflower.

Proof. By Lemma 4.1, there is a w-normal subclass $\mathcal{F}' \subseteq \mathcal{F}$ of size $|\mathcal{F}'| \geq 2^{kw}$. There are two possible cases.

Case 1: There is a subset $S \subseteq X$ such that

$$|\{U \in \mathcal{F}' : S \subseteq U\}| \geq |\mathcal{F}'| \cdot 2^{-k|S|}.$$

Define the family $\mathcal{F}'_S := \{U \setminus S : (U \in \mathcal{F}') \wedge (S \subseteq U)\}$. Notice that

- $\mathcal{F}'_S$ is $(w - |S|)$-normal.
- $|\mathcal{F}'_S| \geq |\mathcal{F}'| \cdot 2^{-k|S|} \geq 2^{k(w-|S|)}$.

By induction both (i) and (ii) hold.

Case 2: For all $S \subseteq X$,

$$|\{U \in \mathcal{F}' : S \subseteq U\}| \leq |\mathcal{F}'| \cdot 2^{-k|S|}$$

Notice that this is the Lipchitz condition for Lemma 4.2 and Lemma 4.5. Their conclusions are precisely (i) and (ii). □

References

[ASU12] Noga Alon, Amir Shpilka, and Christopher Umans. “On sunflowers and matrix multiplication”. In: *Computational Complexity* (2012), pp. 214–223 (cit. on p. 2).

[BADTS17] Avraham Ben-Aroya, Dean Doron, and Amnon Ta-Shma. “Explicit two-source extractors for near-logarithmic min-entropy”. In: *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*. 2017 (cit. on p. 2).

[CFLPP17] Ernie Croot, Vsevolod F. Lev, and Pter Pl Pach. “Progression-free sets in Z_4^n are exponentially small”. In: *Annals of Mathematics* 185.1 (2017), pp. 331–337 (cit. on p. 2).

[CG88] Benny Chor and Oded Goldreich. “Unbiased bits from sources of weak randomness and probabilistic communication complexity”. In: *SIAM Journal on Computing* 17.2 (1988), pp. 230–261 (cit. on p. 7).

[Coh17] Gil Cohen. “Two-Source Extractors for Quasi-Logarithmic Min-Entropy and Improved Privacy Amplification Protocols”. In: *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*. 2017 (cit. on p. 2).

[CZ16] Eshan Chattopadhyay and David Zuckerman. “Explicit Two-Source Extractors and Resilient Functions”. In: *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*. 2016 (cit. on p. 2).

[ER60] Paul Erdős and R Rado. “Intersection theorems for systems of sets”. In: *Journal of the London Mathematical Society* 35.1 (1960), pp. 85–90 (cit. on pp. 1, 2).

[Erd47] Paul Erdős. “Some remarks on the theory of graphs”. In: *Bull. Amer. Math. Soc.* 53.4 (1947), pp. 292–294 (cit. on p. 2).

[ES35] Paul Erdős and George Szekeres. “A combinatorial problem in geometry”. In: *Compositio Mathematica* 2 (1935), pp. 463–470 (cit. on p. 1).

[GLM+16] M. Goos, S. Lovett, R. Meka, T. Watson, and D. Zuckerman. “Rectangles are Nonnegative Juntas”. In: *SIAM Journal on Computing* 45.5 (2016), pp. 1835–1869 (cit. on pp. 3, 5).

[GMR13] Parikshit Gopalan, Raghu Meka, and Omer Reingold. “DNF sparsification and a faster deterministic counting algorithm”. In: *computational complexity* 22.2 (2013), pp. 275–310 (cit. on pp. 2, 3, 5).

[GT08] Ben Green and Terence Tao. “The primes contain arbitrarily long arithmetic progressions”. In: *Annals of Mathematics* 167.2 (2008), pp. 481–547 (cit. on p. 1).

[KMR17] Pravesh K. Kothari, Raghu Meka, and Prasad Raghavendra. “Approximating rectangles by juntas and weakly-exponential lower bounds for LP relaxations of CSPs”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017 (cit. on p. 5).

[Li15] Xin Li. “Three Source Extractors for Polylogarithmic Min-Entropy”. In: *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*. 2015 (cit. on p. 2).

[Li17] Xin Li. “Improved Non-Malleable Extractors, Non-Malleable Codes and Independent Source Extractors”. In: *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*. 2017 (cit. on p. 2).

[Raz85] Alexander Razborov. “Some lower bounds for the monotone complexity of some Boolean functions”. In: *Soviet Math. Dokl.* 31 (1985), pp. 354–357 (cit. on p. 2).

[Ros10] Benjamin Rossman. “The monotone complexity of k-clique on random graphs”. In: *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society. 2010, pp. 193–201 (cit. on p. 2).

[SEG17] Jordan S. Ellenberg and Dion Gijswijt. “On large subsets of F_q^n with no three-term arithmetic progression”. In: *Annals of Mathematics* 185.1 (2017), pp. 339–343 (cit. on p. 2).

[Sze75] Endre Szemerdi. “On sets of integers containing no k elements in arithmetic progression”. In: *Acta Arithmetica* 27 (1975), pp. 199–245 (cit. on p. 1).

[Sze78] Endre Szemerdi. “Regular partitions of graphs”. In: *Problmes combinatoires et thorie des graphes* 260 (1978), pp. 399–401 (cit. on p. 1).

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il