Electronic Colloquium on Computational Complexity, Report No. 173 (2019)

Extractor Lower Bounds, Revisited

Divesh Aggarwal* Siyao Guo† Maciej Obremski‡ João Ribeiro§
Noah Stephens-Davidowitz¶

Abstract

We revisit the fundamental problem of determining seed length lower bounds for strong extractors and natural variants thereof. These variants stem from a "change in quantifiers" over the seeds of the extractor: While a strong extractor requires that the average output bias (over all seeds) is small for all input sources with sufficient min-entropy, a *somewhere* extractor only requires that there *exists* a seed whose output bias is small. More generally, we study what we call *probable* extractors, which on input a source with sufficient min-entropy guarantee that a large enough fraction of seeds have small enough associated output bias. Such extractors have played a key role in many constructions of pseudorandom objects, though they are often defined implicitly and have not been studied extensively.

Prior known techniques fail to yield good seed length lower bounds when applied to the variants above. Our novel approach yields significantly improved lower bounds for somewhere and probable extractors. To complement this, we construct a somewhere extractor that implies our lower bound for such functions is tight in the high min-entropy regime. Surprisingly, this means that a random function is far from an optimal somewhere extractor in this regime. The techniques that we develop also yield an alternative, simpler proof of the celebrated optimal lower bound for strong extractors originally due to Radhakrishnan and Ta-Shma (SIAM J. Discrete Math., 2000).

1 Introduction

Strong seeded extractors are central objects in pseudorandomness that have found many applications in theoretical computer science and cryptography. Informally speaking, a function $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ is a strong extractor if for every source X of sufficiently high min-entropy it holds that the average bias of $\mathsf{Ext}(X,i)$ over the seeds $i \in [D]$ is small. More precisely, we have the definition below. Throughout this paper, we focus on single-bit output extractors since lower bounds in this setting immediately imply lower bounds for any m-bit output extractor.

Definition 1 ((k,ε)-strong extractor)**.** *For $\varepsilon < 1/2$, a function $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ is said to be a (k,ε)-*strong extractor *if*

$$\mathop{\mathbb{E}}_{i \leftarrow [D]} [\Delta(\mathsf{Ext}(X,i); U_1)] \leq \varepsilon \tag{1}$$

*Centre for Quantum Technologies and National University of Singapore. dcsdiva@nus.edu.sg
†New York University Shanghai. siyao.guo@nyu.edu
‡Centre for Quantum Technologies. obremski.math@gmail.com
§Imperial College London. j.lourenco-ribeiro17@imperial.ac.uk
¶Massachusetts Institute of Technology. noahsd@gmail.com. Research supported in part by NSF/BSF grant #1350619, an MIT-IBM grant, and a DARPA Young Faculty Award.

ISSN 1433-8092

for every (n,k)-source[1] X*, where* $i \leftarrow [D]$ *means* i *is uniformly distributed over* $[D]$*, and*

$$\Delta(\mathsf{Ext}(X,i);U_1) = |\Pr[\mathsf{Ext}(X,i)=1] - 1/2|$$

is the bias of $\mathsf{Ext}(X,i)$*.*

A fundamental parameter when studying strong extractors is the number of seeds D. Ideally, one would like to construct strong extractors with D as small as possible. However, there exist lower bounds on D depending on n, k, and ε. Nisan and Zuckerman [NZ96] showed that every strong extractor must use $D = \Omega\big(\frac{n-k}{\varepsilon}\big)$ seeds. Later, in a seminal work, Radhakrishnan and Ta-Shma [RT00] improved the lower bound above to

$$D = \Omega\left(\frac{n-k}{\varepsilon^2}\right). \tag{2}$$

Notably, this lower bound turns out to be tight. In fact, a random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ with $D = C \cdot \frac{n-k}{\varepsilon^2}$ seeds, for a sufficiently large constant $C > 0$, is a (k,ε)-strong extractor with high probability.

At the opposite end of the spectrum lies another well-known pseudorandom object, called a *somewhere extractor*. While a strong extractor has small average bias, all we require of a somewhere extractor is that its *minimum* bias over all seeds is small. More precisely, we have the following definition.

Definition 2 ((k,ε)-somewhere extractor)**.** *For* $\varepsilon < 1/2$*, a function* $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ *is said to be a* (k,ε)-somewhere extractor *if for every* (n,k)*-source* X *it holds that*

$$\min_{i \in [D]} \Delta(\mathsf{Ext}(X,i);U_1) \leq \varepsilon.$$

Somewhere extractors arise in many different contexts, e.g., in the construction of multi-source extractors. Given the complete picture we have of strong extractors, it is natural to wonder what kind of bounds we can prove on the number of seeds D for a somewhere extractor Ext. A simple averaging argument on the preimage sizes of Ext shows that $D > n-k$, but it is possible to improve on this lower bound. If one considers somewhere extractors with m output bits, then [AOR$^+$19] showed that a connection to dispersers leads to the lower bound

$$D = \Omega\left(\frac{n-k}{\varepsilon + 2^{-m}}\right). \tag{3}$$

While (3) was good enough in the context of [AOR$^+$19], it is quite unsatisfactory in general for two reasons: First, it is trivial for small m (e.g., in our setting, where $m = 1$). Second, even for larger m, it does not scale with ε below 2^{-m}. In the 1-bit output setting, which is the hardest for lower bounds, the best known lower bound is [AOR$^+$19]

$$D = \Omega(n - k + \log(1/\varepsilon)).$$

On the other hand, as we show in this work, a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ is a (k,ε)-somewhere extractor with non-negligible probability only when $D = \Omega\big(\frac{1}{\varepsilon^2}\big)$ (we discuss this in more detail below). Therefore, in contrast with strong extractors, there is a large gap between upper and lower bounds on the number of seeds required by somewhere extractors, leaving open the exciting possibility of better constructions.

[1] A distribution $X \in \{0,1\}^n$ is said to be an (n,k)-*source* if it has min-entropy $\mathbf{H}_\infty(X) \geq k$.

One may also wonder whether the strong extractor lower bound techniques from [NZ96, RT00] can be adapted to yield better lower bounds for somewhere extractors. However, it is not clear how this can be done, since these techniques are fundamentally tailored for dealing solely with the average bias as in Definition 1. Overall, current techniques seem incapable of yielding a sharp, unconditional analysis of somewhere extractors.

1.1 Our contributions

In this work, we develop a novel approach towards lower bounding the number of seeds required by natural variants of strong extractors. We highlight our main results here.

1.1.1 Improved lower bounds for somewhere extractors

We significantly improve the lower bound for (k, ε)-somewhere extractors. More precisely, we prove the following result.

Theorem 1. *Every (k, ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ must have*

$$D \geq \frac{\ln 2}{2} \cdot \frac{n-k}{\varepsilon}. \tag{4}$$

Recall that the previous best lower bound was $D = \Omega(n - k + \log(1/\varepsilon))$. Observe also that the lower bound in (4) is a factor of ε smaller than the one in (2) for strong extractors, and is therefore also a factor of ε smaller than the upper bound via the probabilistic method for somewhere extractors. Remarkably, we construct a (k, ε)-somewhere extractor that shows (4) is *tight* in the high min-entropy regime.

Theorem 2. *For every[2] $\varepsilon \geq \frac{1}{2(1+2^k)}$, there exists a (k, ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ with*

$$D = \frac{2^{n-k-1}}{\varepsilon} + 1.$$

In particular, Theorem 2 shows that (4) is tight up to a constant factor when $n - k$ is constant.

On a related front, the existential result for strong extractors immediately implies that a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ with $D = C \cdot \frac{n-k}{\varepsilon^2}$ for a large enough constant $C > 0$ is a (k, ε)-somewhere extractor with high probability. Interestingly, we show that this probabilistic argument is tight up to a constant factor, in the sense that a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ must have $D = \Omega\big(\frac{n-k}{\varepsilon^2}\big)$ in order to be a (k, ε)-somewhere extractor with non-negligible probability for essentially all regimes of k and ε. Given the above, we conclude that a random function is far from an optimal (k, ε)-somewhere extractor in the high min-entropy regime. This is a rare example where an explicit construction is significantly better than the probabilistic method for some regime of parameters. To be more precise, we show the following.

Theorem 3. *For large enough n, suppose that $k \leq n - 100$, $2^{-0.24(n+k)} \leq \varepsilon \leq c_0$ for a sufficiently small constant $c_0 > 0$, and*

$$D \leq \frac{n-k}{100 \cdot \varepsilon^2}.$$

Then, a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ is not *a (k, ε)-somewhere extractor with probability at least $1 - 2^{-2^{\Omega(n)}}$.*

[2] We note that there are no (k, ε)-somewhere extractors with $\varepsilon < \frac{1}{2(1+2^k)}$. To see this, consider an (n, k)-source X uniformly distributed over a set of size $2^k + 1$. Then, $\mathsf{Ext}(X, i)$ has bias at least $\frac{1}{2(1+2^k)}$ for every i.

1.1.2 Simple proof of the optimal lower bound for strong extractors

In the setting of strong extractors, we give an alternative, much simpler proof of the tight lower bound (2) due to Radhakrishnan and Ta-Shma [RT00]. To be precise, we prove the following result.

Theorem 4. *For every $n, k, \varepsilon > 0$ satisfying $n - k \geq 39$ and every (k, ε)-strong extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ it holds that*

$$D \geq \frac{\ln 2}{18} \cdot \frac{n-k}{\varepsilon^2}.$$

1.1.3 Generalizing somewhere extractors and lower bounds

We initiate the systematic study of a meaningful generalization of somewhere extractors and also obtain significantly improved lower bounds in that setting, as discussed below. A somewhere extractor Ext can be generalized in a natural way by requiring that some fraction of the seeds of Ext yield an unbiased output, instead of only a single seed. This leads to the following definition.

Definition 3 ((k, ε, δ)-probable extractor)**.** *For $\varepsilon < 1/2$, a function $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ is said to be a* (k, ε, δ)-probable extractor *if*

$$\Pr_{i \leftarrow [D]}[\Delta(\mathsf{Ext}(X, i); U_1) > \varepsilon] < \delta$$

for every (n, k)-source X.

We note that probable extractors have been defined explicitly before, but not studied in depth, in [Rao07, BCD+18]. Observe that a (k, ε)-somewhere extractor corresponds to a $(k, \varepsilon, \delta = 1)$-probable extractor. Moreover, a (k, ε)-strong extractor lies somewhere between a $(k, \varepsilon/2, \varepsilon/2)$-probable extractor and a $(k, \sqrt{\varepsilon}, \sqrt{\varepsilon})$-probable extractor. More generally, every (k, ε)-strong extractor is a $(k, \varepsilon/\delta, \delta)$-probable extractor for every $\delta > 0$ by Markov's inequality. On the other hand, we also have that every (k, ε, δ)-probable extractor is a $(k, \varepsilon + \delta)$-strong extractor.

Given our previous discussion, a natural question we ask about probable extractors is the following:

How do ε and δ influence the number of seeds D?

Our work leads to a better understanding of this behavior. Similarly to what was already discussed in [BCD+18], by separating the maximum fraction of "bad" seeds δ and the maximum bias of the "good" seeds ε, we are able to explore the explicit influence that each of these important parameters has on the number of seeds. Such a fine-grained analysis is not possible, for example, in the case of strong extractors, since those properties are essentially merged into a single global error parameter.

Besides being interesting on its own, there are practical motivations for the question above. In fact, several constructions of multi-source extractors make use of (k, ε)-strong extractors in scenarios where a $(k, \varepsilon/\delta, \delta)$-probable extractor would suffice with δ much larger than ε. The reason for this is simply that no better constructions of $(k, \varepsilon/\delta, \delta)$-probable extractors are known. However, it could be a priori possible to design a $(k, \varepsilon/\delta, \delta)$-probable extractor requiring much fewer seeds than a (k, ε)-strong extractor. In turn, this would lead to simpler constructions of, and improved parameters for, several multi-source extractors. We expand on this in Section 1.2.

Lower bounds for probable extractors Lower bounds on the number of seeds required by probable extractors can be derived directly from lower bounds for both strong and somewhere extractors. Combining (2) with the fact that every (k, ε, δ)-probable extractor is a $(k, \varepsilon + \delta)$-strong extractor immediately leads to the lower bound

$$D = \Omega\left(\frac{n-k}{(\varepsilon+\delta)^2}\right).$$

However, note that the bound above becomes trivial whenever one of ε or δ is large. Observe that a (k, ε, δ)-probable extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ must be a (k, ε)-somewhere extractor when restricted to the first δD seeds. Therefore, any lower bound L for the number of seeds of (k, ε)-somewhere extractors immediately implies the lower bound $D \geq L/\delta$ for any (k, ε, δ)-probable extractor. Combining this with Theorem 1 leads to the following result.

Theorem 5. *Let* $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ *be a* (k, ε, δ)*-probable extractor. Then, it holds that*

$$D \geq \frac{\ln 2}{2} \cdot \frac{n-k}{\varepsilon \cdot \delta}. \tag{5}$$

The lower bound in (5) significantly improves upon all previous bounds over a large range of (ε, δ), namely when $\delta \gg \varepsilon$ or $\varepsilon \gg \delta$. On the other hand, we show that a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ with $D = O\left(\frac{n-k}{\varepsilon^2 \cdot \delta}\right)$ is a (k, ε, δ)-probable extractor with high probability. It remains an open problem to close the gap between this upper bound and Theorem 5 in general. While we know from our previous discussion that the lower bound in Theorem 5 is tight in some settings, there may be a different behavior when δ is small versus when it is large and for lower min-entropy regimes.

Given the gap between bounds above, it is natural to ask whether a different probabilistic argument could be used to show that a uniformly random function using fewer seeds is a (k, ε, δ)-probable extractor with high probability. As before, we can easily extend Theorem 3 to the setting of probable extractors to show the answer to the question above is negative. Namely, we have the following result, which shows that our probabilistic construction is tight up to a constant factor.

Theorem 6. *For any* $\delta = \delta(n) \in (0, 1]$ *and large enough* n*, suppose that* $k \leq n - 100$*,* $2^{-0.24(n+k)} \leq \varepsilon \leq c_0$ *for a sufficiently small constant* $c_0 > 0$*, and*

$$D \leq \frac{n-k}{100 \cdot \varepsilon^2 \cdot \delta}.$$

Then, a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ *is* not *a* (k, ε, δ)*-probable extractor with probability at least* $1 - 2^{-2^{\Omega(n)}}$*.*

1.2 Applications

Besides the works we have already discussed, several others have either implicitly or explicitly used probable extractors. However, they do not focus on bounding the minimum number of seeds required. Many constructions of seeded and multi-source extractors [Ta-96, LRVW03, Raz05, Rao09, Li11, Li13b, Li13a, Coh15, Li15, Li16, BCD+18, CZ19], along with some constructions of dispersers [BRSW12] and non-malleable and affine extractors [CGL16, Li16], use probable extractors, sometimes satisfying additional properties and taking more objects than just one (n, k)-source as input, as intermediate objects.

In the literature, the output of the probable extractor (concatenated over all D seeds) is usually called a *somewhere-random source with D rows*. The time complexity of the resulting extractor

constructions depends linearly on the complexity of enumerating the D seeds of the probable extractor being used. This poses a problem, because, even now, the best explicit probable extractor we know of for a single weak source is simply a strong extractor. As a result, the lower bound in (2) from [RT00] applies to the number of seeds, and so extra assumptions must be made or parameters must be worsened in order to ensure that seed enumeration can be done efficiently.

We present concrete examples of the compromise above. Some works settle for a large overall $1/\mathrm{poly}(n)$ error of the resulting extractor to get around the seed enumeration problem [Rao09, Li11, Li13b, Li13a, Li16]. On another front, many works use extra independent weak sources with enough min-entropy as input to generate somewhere-random sources with fewer rows [BKS$^+$10, BRSW12, Coh15, Li15, Li16, CGL16, BCD$^+$18, CZ19]. Moreover, the addition of a short uniformly random seed to achieve this goal has also been considered [LRVW03]. Many works above can be interpreted as constructing several types of randomness extractors for somewhere-random sources (called *mergers*), a problem which was first studied by Ta-Shma [Ta-96]. Other works that have studied mergers include [Raz05, Zuc06, DS07, DR08, DW11, DKSS13].

Prior to this work, we could not rule out a (k, ε, δ)-probable extractor for δ much larger than ε with much fewer seeds than a (k, ε)-strong extractor. Given the discussion above, this would lead not only to extractors with improved parameters, but also to conceptually simpler constructions, since many tools and assumptions were introduced to deal with the fact that somewhere-random sources generated by strong extractors have too many rows. Our results preclude this possibility.

Finally, we note that many of the applications above still work if one considers an extractor that outputs *convex combinations* of somewhere-random sources from (n, k)-sources instead. Our lower bounds do not apply to this weaker setting. Therefore, we do not rule out the existence of methods of generating a convex combination of somewhere-random sources from one weak source requiring fewer seeds. Our results show that, without considering convex combinations, one cannot do *too much* better than the naive and globally used method of enumerating over the seeds of a strong extractor (although, surprisingly, we show that a polynomial improvement in $1/\varepsilon$ *is* possible). We leave it as an interesting open problem to extend our new techniques and bounds to the setting of convex combinations.

1.3 Technical overview

In this section, we provide a more detailed account of our contributions.

1.3.1 The high-level approach

Our extractor lower bounds can be unified under a common high-level approach. Fix an arbitrary function $F : \{0,1\}^n \times [D] \to \{0,1\}$. Our goal is to relate the number of seeds D to some measure of the bias of F over all seeds, depending on the type of extractor we are dealing with. For the remainder of this section, we focus on somewhere extractors and the minimum bias. However, everything is equally applicable to strong extractors simply by replacing minimum bias by average bias.

In order to prove a lower bound on D, we follow the high-level recipe below. This corresponds to the natural strategy for deriving lower bounds on D, i.e., by proving the existence of an input (n, k)-source X such that $P = (F(X,1), F(X,2), \ldots, F(X,D))$ is sufficiently biased. However, it is crucially rephrased in a way that allows us to focus directly on output distributions P, instead of defining them indirectly by first designing a suitable input distribution. Throughout the remainder of this paper, for the sake of simplicity we shall write

$$F(x) = (F(x,1), F(x,2), \ldots, F(x,D)) \in \{0,1\}^D.$$

For F and fixed k, we construct a distribution $P = (P_1, P_2, \ldots, P_D)$ over $\{0,1\}^D$ such that:

1. There exists an (n,k)-source X such that $F(X) = P$;
2. It holds that[3]
$$\min_{i \in [D]} \Delta(P_i; U_1) \geq \alpha,$$
where α is some quantity depending on n, k, and D.

If F is a (k,ε)-somewhere extractor, the two conditions above imply that $\varepsilon \geq \alpha$. This relationship then yields a lower bound on D.

The main novelty of our approach lies in the design of the output distribution P. The distribution $A = F(U_n)$ takes on a special role in our construction of good choices of P. We begin by showing that the first condition above automatically holds provided P satisfies a simple constraint related to A, which is detailed in the following lemma (below and throughout the paper, we write $X(x)$ for the probability that a random variable/distribution X takes on value x).

Lemma 1. *There exists an (n,k)-source X such that $F(X) = P$ if*
$$P(a) \leq 2^{n-k} A(a) \tag{6}$$
for all $a \in \{0,1\}^D$.

Proof. It is enough to consider the source $X \in \{0,1\}^n$ that picks each $x \in \{0,1\}^n$ with probability
$$X(x) = 2^{-n} \cdot \frac{P(F(x))}{A(F(x))}.$$
First, by (6) it follows that $X(x) \leq 2^{-n} \cdot 2^{n-k} = 2^{-k}$ for every x. Moreover, using the fact that $A(a) = 2^{-n} \cdot |F^{-1}(a)|$, it is easy to see that X is a valid probability distribution and $F(X) = P$. □

Remark 1. It is easy to see that the statement in Lemma 1 is actually an "if and only if". However, in this work we only require the "if" part.

We construct distributions P implicitly in terms of the distribution $A = F(U_n)$. In fact, Lemma 1 shows it is enough to restrict our attention to distributions P that can be written as
$$P(a) = A(a) \cdot f(a)$$
for some non-negative function f satisfying $f(a) \leq 2^{n-k}$ for all $a \in \{0,1\}^D$. As discussed in the following sections, careful choices of f lead to good lower bounds on D with streamlined derivations.

1.3.2 Improved lower bound for somewhere and probable extractors

We employ the high-level approach detailed in Section 1.3.1 to obtain improved lower bounds for probable extractors. Namely, we prove Theorem 5, which states that every (k,ε,δ)-probable extractor must have $D \geq \frac{\ln 2}{2} \cdot \frac{n-k}{\varepsilon \cdot \delta}$.

As discussed in Section 1.1.3, Theorem 5 is a simple corollary of Theorem 1. With this in mind, from here onwards we focus solely on proving this result, which states the lower bound $D \geq \frac{\ln 2}{2} \cdot \frac{n-k}{\varepsilon}$ for (k,ε)-somewhere extractors.

[3] In the case of strong extractors, this condition is replaced by $\mathbb{E}_{i \leftarrow [D]}[\Delta(P_i; U_1)] \geq \alpha$.

Let $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ be an arbitrary (k,ε)-somewhere extractor. In order to prove Theorem 1 via the high-level approach from Section 1.3.1, we consider the family of distributions $\overline{P}_z$ parameterized by $z \in \{0,1\}^D$ defined as

$$\overline{P}_z(a) = \frac{1}{C_z} \cdot A(a) \prod_{i=1}^{D} [1 + (-1)^{a_i + z_i} \gamma], \quad a \in \{0,1\}^D,$$

where $A = \mathsf{Ext}(U_n)$, $\gamma \in (0,1)$ is a parameter of our choice, and C_z is the normalizing factor.

We choose $z^\star$ which maximizes C_z over all $z \in \{0,1\}^D$, and consider $P = \overline{P}_{z^\star}$. In particular, this choice implies that $C_{z^\star} \geq 1$, which allows us to take $\gamma = \Theta\big(\frac{n-k}{D}\big)$ while still satisfying (6). It remains to lower bound $\Delta(P_i; U_1)$ for every $i \in [D]$ appropriately. The product structure of the family of distributions we consider makes it amenable to a Fourier-analytic approach, which we employ to show that for every $i \in [D]$ we have

$$\Delta(P_i; U_1) = \Omega(\gamma) = \Omega\left(\frac{n-k}{D}\right).$$

This yields the desired lower bound on D. More details can be found in Section 3.1.

1.3.3 Tight upper bound for somewhere extractors

We design a somewhere extractor that shows our lower bound for (k,ε)-somewhere extractors is tight (up to a multiplicative constant) in the high min-entropy regime where $n - k = O(1)$. More precisely, we prove Theorem 2, which states that there exists a (k,ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ with $D = \frac{2^{n-k-1}}{\varepsilon} + 1$ for all non-trivial ε.

This is accomplished by showing that the function $\mathsf{Ext} : [N] \times \{0,1,\ldots,E\} \to \{0,1\}$, with $E = \frac{2^{n-k-1}}{\varepsilon}$ and $N = 2^n$, defined as

$$\mathsf{Ext}(x,i) = \mathsf{sign}[(x+i) \mod 2E], \quad i = 0,1,\ldots,E \tag{7}$$

is a (k,ε)-somewhere extractor. In (7), we see $x \mod 2E$ as an integer in $\{-E,\ldots,E-1\}$, and define $\mathsf{sign}(y) = \mathbf{1}_{\{y \geq 0\}}$. Intuitively, this simple function yields a good somewhere extractor because the functions $\mathsf{Ext}(\cdot,i)$ "transition smoothly" from $\mathsf{Ext}(\cdot,0)$ to its opposite, $\mathsf{Ext}(\cdot,E) = 1 - \mathsf{Ext}(\cdot,0)$, as shown in Figure 1.

Given an (n,k)-source X, we wish to prove that there is a seed i such that $\Delta(\mathsf{Ext}(X,i); U_1) \leq \varepsilon$. In order to show this, we will look at how the quantities

$$\Delta_i = \Pr[\mathsf{Ext}(X,i) = 1] - \Pr[\mathsf{Ext}(X,i) = 0], \quad i = 0,1,\ldots,E$$

behave. The desired result follows if we show that $|\Delta_i| \leq 2\varepsilon$ for some i. In turn, this holds because the Δ_i's satisfy two simple properties. First, we have $\Delta_0 = -\Delta_E$. Second, when going from $\mathsf{Ext}(\cdot,i-1)$ to $\mathsf{Ext}(\cdot,i)$, by our choice of parameters at most $4\varepsilon \cdot 2^k$ elements of $[N]$ go from 1 to 0, and vice-versa. This implies that $|\Delta_i - \Delta_{i-1}| \leq 4\varepsilon$. Combining the two properties above immediately ensures the existence of $i^\star$ such that $|\Delta_{i^\star}| \leq 2\varepsilon$, as desired. For more details, see Section 4.

1.3.4 Simpler proof of the optimal lower bound for strong extractors

In this section, we discuss our alternative, simpler proof of the optimal lower bound on the number of seeds for strong extractors, originally obtained by Radhakrishnan and Ta-Shma [RT00]. Namely,

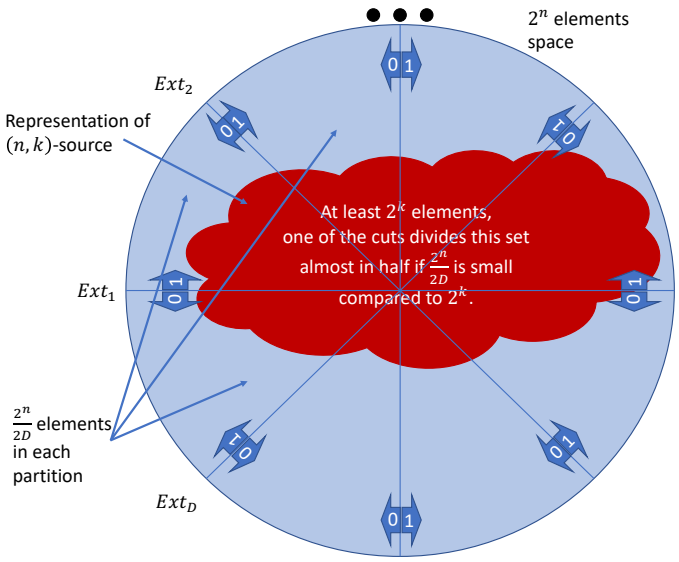

Figure 1: An illustration of the (k, ε)-somewhere extractor that shows our lower bound is tight in the high min-entropy regime.

we prove Theorem 4, which states that every (k, ε)-strong extractor must have $D \geq \frac{\ln 2}{18} \cdot \frac{n-k}{\varepsilon^2}$ when $n - k \geq 39$.

As before, we follow the high-level approach introduced in Section 1.3.1. However, we consider a different family of distributions. Fix an arbitrary (k, ε)-strong extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$, and let $A = \mathsf{Ext}(U_n)$. Then, for $z \in \{0,1\}^D$ and $t \leq D$ define the distribution

$$\overline{P}_{z,t}(a) = \frac{1}{C_{z,t}} A(a) \cdot \mathbf{1}_{\{\|z-a\|_1 \leq t\}}, \quad a \in \{0,1\}^D,$$

where $C_{z,t}$ is the normalizing factor. The desired result now follows via two simple combinatorial arguments, which guarantee that (i) for an appropriate $t = D/2 - \Theta(\sqrt{(n-k)D})$, there exists a choice of z such that $P = \overline{P}_{z,t}$ satisfies (6), and (ii) the average bias of every distribution $\overline{P}_{z,t}$ is at least $1/2 - t/D = \Omega\left(\sqrt{(n-k)/D}\right)$. More details can be found in Section 3.2.

1.3.5 Probabilistic constructions and lower bounds for random functions

We study for which values of D it holds that a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ is a (k, ε, δ)-probable extractor with non-negligible probability. To show an upper bound, we consider a connection between probable extractors and strong *two-source* extractors, and then invoke well-known existential results for the latter.[4] This shows that, under a mild constraint on k, ε, and δ, a uniformly random function is a (k, ε, δ)-probable extractor with probability at least, say, 0.99 when $D = C \cdot \frac{n-k}{\varepsilon^2 \cdot \delta}$ for a sufficiently large constant $C > 0$.

We complement the upper bound in the previous paragraph via Theorem 6, which states that a uniformly random function with $D \leq \frac{n-k}{100 \cdot \varepsilon^2 \cdot \delta}$ is *not* a (k, ε, δ)-probable extractor with probability

[4]More direct approaches do not seem to work because the set of sources from which a somewhere extractor successfully extracts is not necessarily convex (see Remark 2 in Section 5).

at least $1 - 2^{-2^{\Omega(n)}}$. This means that our probabilistic construction above is tight up to a constant factor. Similarly to Section 1.3.2, to prove this result it suffices to focus our attention on somewhere extractors. We consider a source $X \in \{0,1\}^n$ uniformly distributed over a set $\mathcal{X}_F$ defined as

$$\mathcal{X}_F = \{x \in \{0,1\}^n : \|F(x)\|_1 \leq t\}$$

for an appropriate $t = D/2 - \Theta(\sqrt{(n-k)D})$. Then, we show X satisfies two properties: First, by a Chernoff bound, it holds that $|\mathcal{X}_F| \geq 2^k$ with very high probability over the choice of F, and hence X is an (n,k)-source with very high probability. Second, we show that, again with very high probability, we have $\Delta(F(X,i);U_1) > \varepsilon$ simultaneously for all $i \in [D]$. These two properties immediately imply that F is not a (k,ε)-somewhere extractor with high probability. More details can be found in Section 5.

1.4 Open questions

Besides the natural problem of improving upon our lower bounds in general, our work leaves open other interesting avenues for further research.:

- Consider the special case of $(k,\varepsilon,\varepsilon)$-probable extractors. In this setting, the best lower bounds only yield $D = \Omega\big(\frac{n-k}{\varepsilon^2}\big)$, while the probabilistic method requires $D = \Omega\big(\frac{1}{\varepsilon^3}\big)$ to work with non-negligible probability. We believe $D = \Theta\big(\frac{n-k}{\varepsilon^3}\big)$ is the correct answer, and it would be very interesting to prove (or disprove) this claim, as it showcases different behavior than (k,ε,δ)-probable extractors for $\delta \gg \varepsilon$;
- Show the existence of a (k,ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ using $D = O\big(\frac{n-k}{\varepsilon}\big)$ seeds. This would extend the tightness of our lower bound for somewhere extractors below the high min-entropy regime;
- Extend our (k,ε)-somewhere extractor from Section 1.3.3 to output $m > 1$ bits with (roughly) the same number of seeds;
- Extend our lower bounds to the setting where one is allowed to output convex combinations of somewhere-random sources from one (n,k)-source (see Section 1.2).

1.5 Organization

We introduce basic notions and results that are useful throughout our work in Section 2. The proofs of our extractor lower bounds are presented in Section 3. The matching upper bound on the number of seeds of somewhere extractors can be found in Section 4. Finally, probabilistic constructions of probable extractors, along with lower bounds on the number of seeds of uniformly random functions are discussed in Section 5.

2 Preliminaries

2.1 Notation

Random variables and distributions are usually denoted by uppercase letters such as X, Y, and Z. When context allows, we may confuse a random variable with its associated distribution. We write $X(x)$ for the probability that X equals x, and denote the support of X by $\mathsf{supp}(X)$. The uniform distribution over $\{0,1\}^n$ is denoted by U_n. We write $i \leftarrow S$ to mean that i is sampled

uniformly at random from the set S. For a distribution X, we write $x \sim X$ to denote x is sampled according to X. Given an event E, the indicator of E is denoted by $\mathbf{1}_{\{E\}}$. The expected value of a random variable X is denoted by $\mathbb{E}[X]$ or $\mathbb{E}_{x\sim X}[x]$. Sets are usually denoted by uppercase letters such as S and T. The set $\{1, 2, \ldots, D\}$ is denoted by $[D]$. We will usually identify a set S with its characteristic vector, so that we write $S_i = 1$ if and only if $i \in S$. We write $S+T$ for the symmetric difference between two sets S and T (i.e., the modulo 2 sum of their characteristic vectors). We denote the base-2 logarithm by log and the natural logarithm by ln. We write $\|x\|_p$ for the p-norm of a vector x. The inner product between two vectors x and y over some field is denoted by $x \cdot y$.

2.2 Probability theory

In this section, we introduce some basic notions and results from probability theory.

Definition 4 (Statistical distance)**.** *Given two distributions X and Y over a set $\mathcal{X}$, the* statistical distance between X and Y*, denoted by $\Delta(X;Y)$, is defined as*

$$\Delta(X;Y) = \max_{S\subseteq\mathcal{X}} |\Pr[X \in S] - \Pr[Y \in S]| = \frac{1}{2}\sum_{x\in\mathcal{X}} |X(x) - Y(x)|.$$

*We say that X and Y are ε-*close*, also written $X \approx_\varepsilon Y$, if $\Delta(X;Y) \le \varepsilon$.*

Definition 5 (Min-entropy)**.** *Given a distribution X over $\mathcal{X}$, the* min-entropy of X*, denoted by $\mathbf{H}_\infty(X)$, is defined as*

$$\mathbf{H}_\infty(X) = -\log\left(\max_{x\in\mathcal{X}} X(x)\right).$$

Definition 6 ((n,k)-source)**.** *A distribution X supported on $\{0,1\}^n$ is said to be an (n,k)-*source *if $\mathbf{H}_\infty(X) \ge k$. An (n,k)-source is said to be* flat *if it is uniformly distributed over a subset of $\{0,1\}^n$ of size 2^k.*

The several notions of extractors that we focus on in this work were already covered in Definitions 1, 2, and 3 in Section 1.

Later on, we will exploit the well-known (and not difficult to prove) fact that the Chernoff bound is tight (up to constants in the exponent).

Lemma 2 (Inverse Chernoff bound, see, e.g., [KY15, Lemma 4, Part 1 with $p = 1/2$])**.** *Suppose $\gamma, D > 0$ are such that $\gamma \le 1/2$ and $\gamma^2 D \ge 6$, and let Z denote a binomial distribution with D trials and success probability $1/2$. Then,*

$$2^{-D} \cdot \sum_{i=0}^{(1-\gamma)D/2} \binom{D}{i} = \Pr\left[Z \le \frac{(1-\gamma)D}{2}\right] \ge \exp\left(-\frac{9\gamma^2 D}{2}\right).$$

2.3 Basic boolean functional analysis

In this section, we briefly discuss basic notions from the analysis of boolean functions that we will use later on.

Given a set $S \subseteq [n]$, the Fourier character $\chi_S : \{0,1\}^n \to \{-1,1\}$ is defined as

$$\chi_S(x) = (-1)^{x\cdot s},$$

where s is the characteristic vector of S (i.e., $s_i = 1$ if and only if $i \in S$). The characters χ_S satisfy $\chi_S(x+y) = \chi_S(x) \cdot \chi_S(y)$ and form an orthonormal basis of the space of functions $f : \{0,1\}^n \to \mathbb{R}$. Consequently, every such function f has a unique Fourier expansion

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \cdot \chi_S(x),$$

where $\widehat{f}(S) = \mathbb{E}_{x \leftarrow \{0,1\}^n}[f(x) \cdot \chi_S(x)]$ is the Fourier coefficient of f on S.

3 Extractor lower bounds

3.1 A lower bound for probable extractors

In this section, we follow the high-level approach described in Section 1.3.1 to prove Theorem 1, which we restate here for convenience. By the discussion in Section 1.3.2, this result immediately implies the more general Theorem 5 for probable extractors.

Theorem (Theorem 1, restated)**.** *Every (k, ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ must have*

$$D \geq \frac{\ln 2}{2} \cdot \frac{n-k}{\varepsilon}.$$

Fix a (k, ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$, and let $A = \mathsf{Ext}(U_n)$. For $z \in \{0,1\}^D$, consider the (unnormalized) distribution P_z defined as

$$P_z(a) = A(a) \cdot \prod_{i=1}^{D} [1 + (-1)^{a_i + z_i} \gamma] = A(a) \sum_{S \subseteq [D]} \chi_S(a+z) \gamma^{|S|} \tag{8}$$

for $a \in \{0,1\}^D$, where $\gamma = (g/D) \ln 2$ and $g = n - k$ is the min-entropy gap (we shall see that $\varepsilon = \Omega(\gamma)$, which leads to the desired lower bound with this choice of γ). The second equality in (8) holds because for every $S \subseteq [D]$ we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{a \leftarrow \{0,1\}^D} \left[\prod_{i=1}^{D} [1 + (-1)^{a_i + z_i} \gamma] \cdot \chi_S(a) \right] &= \prod_{i=1}^{D} \mathop{\mathbb{E}}_{a_i \leftarrow \{0,1\}} [(1 + (-1)^{a_i + z_i} \gamma) \cdot (-1)^{a_i \cdot S_i}] \\
&= \prod_{i \in S} \gamma (-1)^{z_i} \\
&= \gamma^{|S|} \cdot \chi_S(z).
\end{aligned}$$

Observe that $0 < \gamma < 1$ since we know $D > g > 0$. To see this, note that, if $D \leq g$, a simple averaging argument guarantees there is $a \in \{0,1\}^D$ with $|\mathsf{Ext}^{-1}(a)| \geq 2^k$. This implies there is an (n,k)-source X such that $\mathsf{Ext}(X)$ is constant. We can then define the normalizing constant

$$C_z = \sum_{a \in \{0,1\}^D} P_z(a) > 0,$$

and we set

$$\overline{P}_z(a) = \frac{P_z(a)}{C_z}.$$

We will fix $z^\star$ to be the choice of z that maximizes C_z, and we let

$$P = \overline{P}_{z^\star}.$$

Note that P is a distribution over $\{0,1\}^D$. We denote the distribution of its i-th coordinate by P_i.

Proof of Theorem 1 Our goal now is twofold: First, we must ensure that

$$P(a) \leq 2^g A(a) \quad \forall a \in \{0,1\}^D. \tag{9}$$

Second, we wish to show that

$$\min_{i \in [D]} \Delta(P_i; U_1) \geq \gamma/2. \tag{10}$$

Since Ext is a (k, ε)-somewhere extractor, from (9) and (10) it follows that $\varepsilon \geq \gamma/2$. By the choice of γ above this immediately implies Theorem 1.

Lemma 3. *Condition* (9) *holds for the choice of* $z^\star$ *and* γ *above.*

Proof. First, since $z^\star = \arg\max_z C_z$, we have

$$\begin{aligned} C_{z^\star} &\geq \mathop{\mathbb{E}}_{z \leftarrow \{0,1\}^D}[C_z] \\ &= \sum_{a \in \{0,1\}^D} \mathop{\mathbb{E}}_{z \leftarrow \{0,1\}^D}[P_z(a)] \\ &= \sum_{a \in \{0,1\}^D} A(a) \\ &= 1. \end{aligned}$$

Therefore, it suffices to show that

$$\prod_{i=1}^{D} [1 + (-1)^{z_i^\star + a_i} \gamma] \leq 2^g.$$

for every $a \in \{0,1\}^D$. This follows immediately from the fact that

$$(1+\gamma)^D = \left(1 + \frac{g \ln 2}{D}\right)^D \leq \exp(g \ln 2) = 2^g. \qquad \square$$

Lemma 4. *We have*

$$\min_{i \in [D]} \Delta(P_i; U_d) \geq \gamma/2.$$

Proof. In order to show the desired inequality, it suffices to prove that

$$\left| \mathop{\mathbb{E}}_{a \sim P}[\chi_{\{i\}}(a)] \right| \geq \gamma \tag{11}$$

for every $i \in [D]$. For any $T \subseteq [D]$ and $z \in \{0,1\}^D$, we have

$$\begin{aligned} C_z \cdot \mathop{\mathbb{E}}_{a \sim \overline{P}_z}[\chi_T(a)] &= \sum_{a \in \{0,1\}^D} P_z(a) \cdot \chi_T(a) \\ &= \sum_{a \in \{0,1\}^D} \left[A(a) \sum_{S \subseteq [D]} \chi_S(a+z) \gamma^{|S|} \right] \cdot \chi_T(a) \\ &= \sum_{S \subseteq [D]} \chi_S(z) \gamma^{|S|} \sum_{a \in \{0,1\}^D} A(a) \cdot \chi_S(a) \cdot \chi_T(a) \end{aligned}$$

$$= 2^D \sum_{S\subseteq[D]} \chi_S(z)\gamma^{|S|} \cdot \widehat{A}(S+T), \tag{12}$$

where the second equality follows from (8), the third equality is true because $\chi_S(a+z) = \chi_S(a) + \chi_S(z)$, and the last equality holds since $\chi_S(a) \cdot \chi_T(a) = \chi_{S+T}(a)$ and by the definition of $\widehat{A}(S+T)$.

For $i \in [D]$ and $b \in \{0,1\}$, define

$$a_b = 2^D \sum_{S\subseteq[D]:S_i=b} \chi_S(z^\star)\gamma^{|S|} \cdot \widehat{A}(S).$$

By setting $T = \emptyset$ and $z = z^\star$, from (12) we obtain

$$C_{z^\star} = C_{z^\star} \cdot \mathop{\mathbb{E}}_{a\sim P}[\chi_\emptyset(a)] = a_0 + a_1. \tag{13}$$

Moreover, setting $T = \{i\}$ and $z = z^\star$ in (12) leads to

$$\begin{aligned}
C_{z^\star} \cdot \mathop{\mathbb{E}}_{a\sim P}[\chi_{\{i\}}(a)] &= 2^D \sum_{S\subseteq[D]} \chi_S(z^\star)\gamma^{|S|} \cdot \widehat{A}(S+\{i\}) \\
&= 2^D \sum_{S':=S+\{i\}\subseteq[D]} \chi_{S'+\{i\}}(z^\star)\gamma^{|S'+\{i\}|} \cdot \widehat{A}(S') \\
&= \chi_{\{i\}}(z^\star) \cdot \gamma \cdot 2^D \sum_{S'\subseteq[D]:S'_i=0} \chi_{S'}(z^\star) \cdot \gamma^{|S'|}\widehat{A}(S') \\
&\quad + \frac{\chi_{\{i\}}(z^\star)}{\gamma} \cdot 2^D \sum_{S'\subseteq[D]:S'_i=1} \chi_{S'}(z^\star) \cdot \gamma^{|S'|}\widehat{A}(S') \\
&= \chi_{\{i\}}(z^\star)\left(a_0\gamma + \frac{a_1}{\gamma}\right).
\end{aligned} \tag{14}$$

Combining (14) with (13) implies that

$$\left|\mathop{\mathbb{E}}_{a\sim P}[\chi_{\{i\}}(a)]\right| = \left|\frac{a_0\gamma + a_1/\gamma}{a_0 + a_1}\right|. \tag{15}$$

To conclude the proof, we show that $a_1 \geq 0$. Coupled with (15), this yields (11) because then we have

$$|a_0\gamma + a_1/\gamma| \geq \gamma(a_0 + a_1) = \gamma|a_0 + a_1|,$$

where the inequality follows from $a_1 \geq 0$ and the fact that $0 < \gamma < 1$ (recall that $D > g > 0$), and the equality holds because $a_0 + a_1 = C_{z^\star} > 0$.

It remains to show that $a_1 \geq 0$. Let $e_i \in \{0,1\}^D$ be the vector that is 1 at i and 0 elsewhere, and set $z' = z^\star + e_i$. Then, by (12) with $T = \emptyset$ and $z = z' = z^\star + e_i$ we have

$$\begin{aligned}
C_{z'} &= 2^D \sum_{b\in\{0,1\}} \sum_{S\subseteq[D]:S_i=b} \chi_S(z^\star + e_i)\gamma^{|S|} \cdot \widehat{A}(S) \\
&= 2^D \sum_{b\in\{0,1\}} (-1)^b \sum_{S\subseteq[D]:S_i=b} \chi_S(z^\star)\gamma^{|S|} \cdot \widehat{A}(S) \\
&= a_0 - a_1,
\end{aligned}$$

where the second equality follows from the multiplicative property of χ_S and the fact that $\chi_S(e_i) = (-1)^{S_i}$. Since $z^\star = \arg\max_z C_z$, we conclude that $a_0 + a_1 = C_{z^\star} \geq C_{z'} = a_0 - a_1$, and thus $a_1 \geq 0$. □

3.2 A lower bound for strong extractors

In this section, we prove Theorem 4 via the high-level approach in Section 1.3.1. This yields a different, simpler proof of the optimal lower bound on the number of seeds of (k, ε)-strong extractors, originally obtained by Radhakrishnan and Ta-Shma [RT00]. We restate Theorem 4 here for convenience.

Theorem (Theorem 4, restated)**.** *For every $n, k, \varepsilon > 0$ satisfying $n - k \geq 39$ and every (k, ε)-strong extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ it holds that*

$$D \geq \frac{\ln 2}{18} \cdot \frac{n-k}{\varepsilon^2}.$$

Fix a (k, ε)-strong extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$, and let $A = \mathsf{Ext}(U_n)$. For $z \in \{0,1\}^D$ and $t \leq D$, define the (unnormalized) distribution $P_{z,t}$ over $\{0,1\}^D$ as

$$P_{z,t}(a) = A(a) \cdot \mathbf{1}_{\{\|z-a\|_1 \leq t\}}.$$

The associated normalizing factor $C_{z,t}$ is given by

$$C_{z,t} = \sum_{a \in \{0,1\}^D} P_{z,t}(a) = A(B_t(z)),$$

where $B_t(z)$ denotes the Hamming ball of radius t centered at z, and $A(B_t(z)) = \sum_{a \in B_t(z)} A(a)$ denotes its measure under A. Provided that $A(B_t(z)) > 0$, we can then define the normalized distribution $\overline{P}_{z,t}$ by

$$\overline{P}_{z,t}(a) = \frac{P_{z,t}(a)}{C_{t,z}}.$$

Proof of Theorem 4 Taking into account Section 1.3.1, in order to prove Theorem 4 we will show, via two easy lemmas, that $P = \overline{P}_{z^\star,t^\star}$ for appropriate choices $z^\star$ and $t^\star$ satisfies

$$P(a) \leq 2^g A(a) \quad \forall a \in \{0,1\}^D \tag{16}$$

and

$$\mathop{\mathbb{E}}_{i \leftarrow [D]}[\Delta(P_i; U_1)] = \frac{1}{D} \left\| \mathop{\mathbb{E}}_{a \sim P}[a] - (1/2, \ldots, 1/2) \right\|_1 \geq c\sqrt{g/D}, \tag{17}$$

where $g = n - k$ is the entropy gap, $c = \sqrt{\frac{\ln 2}{18}}$, and P_i denotes the distribution of the i-th coordinate of P. Properties (16) and (17) imply there is an (n, k)-source X such that $\mathsf{Ext}(X) = P$, and so $\varepsilon \geq c\sqrt{g/D}$. This immediately yields the desired lower bound on D.

Lemma 5. *For $c = \sqrt{\frac{\ln 2}{18}}$ and $C = 39$, if $g = n - k \geq C$, there exists $z^\star \in \{0,1\}^D$ such that (16) is satisfied for $t^\star = D/2 - c\sqrt{gD}$.*

Proof. Note that (16) is equivalent to

$$A(B_{t^\star}(z^\star)) \geq 2^{-g}.$$

Moreover, a simple averaging argument (based on the fact that every $y \in \{0,1\}^D$ belongs to the same number of Hamming balls) implies there is $z^\star$ such that $A(B_{t^\star}(z^\star)) \geq 2^{-D} \cdot V_{t^\star}$ (recall $V_{t^\star}$ denotes the volume of a Hamming ball of radius $t^\star$). Fix this choice of $z^\star$. From the choice of c

and C above, and since $D > g \geq C$, by the inverse Chernoff bound (Lemma 2 with $\gamma = 2c\sqrt{g/D}$) we have

$$2^{-D} \cdot V_{t^\star} \geq \exp\bigl(-18c^2 g\bigr) = 2^{-g}$$

for $t^\star = D/2 - c\sqrt{gD}$. □

By considering the shifted extractor $\overline{\mathsf{Ext}}(x) = \mathsf{Ext}(x) + z^\star$, without loss of generality we can assume that $z^\star = 0$. Then, we have the following result.

Lemma 6. *For $t = D/2 - \alpha$, it holds that*

$$\left\| \mathop{\mathbb{E}}_{a \sim \overline{P}_{0,t}} [a] - (1/2, \ldots, 1/2) \right\|_1 \geq \alpha.$$

Proof. Note that $\mathbb{E}_{a \sim \overline{P}_{0,t}}[a]$ is a convex combination of elements of $\mathsf{supp}(\overline{P}_{0,t})$. Since $\mathsf{supp}(\overline{P}_{0,t}) \subseteq B_t(0)$, it follows that

$$\left\| \mathop{\mathbb{E}}_{a \sim \overline{P}_{0,t}} [a] \right\|_1 \leq t.$$

Moreover, it holds that $\|(1/2, \ldots, 1/2)\|_1 = D/2$. Consequently, by the triangle inequality we have

$$\left\| \mathop{\mathbb{E}}_{a \sim \overline{P}_{0,t}} [a] - (1/2, \ldots, 1/2) \right\|_1 \geq D/2 - t = \alpha.$$ □

Combining Lemmas 5 and 6 immediately yields (16) and (17).

4 Matching upper bound for somewhere extractors

In this section, we prove Theorem 2, which we restate here.

Theorem (Theorem 2, restated)**.** *There exists a (k, ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ with*

$$D = \frac{2^{n-k-1}}{\varepsilon} + 1.$$

Combining this result with Theorem 1 in the high min-entropy regime (i.e., $n - k = O(1)$) immediately leads to the following corollary.

Corollary 1. *The minimum number of seeds required for a (k, ε)-somewhere extractor $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ when $n - k = O(1)$ is $D = \Theta(1/\varepsilon)$.*

It is instructive to compare Corollary 1 with analogous results for strong extractors and dispersers, since somewhere extractors lie between the two. With respect to dispersers, the optimal number of seeds in the high min-entropy regime is also $\Theta(1/\varepsilon)$ [RT00]. Moreover, this is achieved by a uniformly random function with high probability. For strong extractors, the optimal number of seeds is $\Theta(1/\varepsilon^2)$, again achieved by a uniformly random function with high probability. Remarkably, by Corollary 1 the optimal number of seeds for (1-bit output) somewhere extractors is $\Theta(1/\varepsilon)$, matching the behavior of dispersers. On the other hand, by Theorem 3, a uniformly random function requires $D = \Theta(1/\varepsilon^2)$ to be a (k, ε)-somewhere extractor with non-negligible probability, similarly to strong extractors!

We now proceed to define and analyze the relevant (k,ε)-somewhere extractor Ext that proves Theorem 2. In this section, it will be useful to identify the set of inputs $\{0,1\}^n$ with the set of integers $[N]$ for $N = 2^n$. For any N and D, we define the function $\mathsf{Ext} : [N] \times \{0,1,\ldots,E\} \to \{0,1\}$ for $E = \frac{2^{n-k-1}}{\varepsilon}$ via the simple expression

$$\mathsf{Ext}(x,i) = \mathsf{sign}[(x+i) \mod 2E]. \tag{18}$$

In (18), we interpret $(x+i) \mod 2E$ as an integer in $\{-E,-E+1,\ldots,E-1\}$ and $\mathsf{sign}(y) = \mathbf{1}_{\{y \geq 0\}}$.

Proof of Theorem 2 In order to prove the desired statement for the choice of Ext above, we need to show that for every (n,k)-source X there is $i = 0,1,\ldots,E$ such that

$$\Delta(\mathsf{Ext}(X,i);U_1) = \frac{1}{2}|\Pr[\mathsf{Ext}(X,i)=1] - \Pr[\mathsf{Ext}(X,i)=0]| < \varepsilon.$$

Fix an arbitrary (n,k)-source X. For each seed $i = 0,1,\ldots,E$, define

$$\Delta_i = \Pr[\mathsf{Ext}(X,i)=1] - \Pr[\mathsf{Ext}(X,i)=0].$$

The Δ_i's satisfy two important properties. First, observe that

$$\Delta_0 = -\Delta_E \tag{19}$$

since

$$\mathsf{sign}(x \mod 2D) = 1 - \mathsf{sign}[(x+E) \mod 2E]$$

for every x. Second, for every $i \in [E]$ it holds that

$$|\Delta_i - \Delta_{i-1}| \leq 2^{-k} \cdot \left\lceil \frac{N}{2E} \right\rceil = 2^{-k} \cdot \left\lceil \varepsilon 2^k \right\rceil \leq 4\varepsilon. \tag{20}$$

To see that (20) holds, it suffices to note that (i) there are at most $\lceil \frac{N}{2E} \rceil = \lceil \varepsilon 2^k \rceil$ integers $x \in [N]$ such that $\mathsf{Ext}(x,i-1) = 0$ but $\mathsf{Ext}(x,i) = 1$ and vice-versa, (ii) $X(x) \leq 2^{-k}$ for every integer x, and (iii) we have $\varepsilon \geq \frac{1}{2(1+2^k)}$. Finally, combining (19) and (20) ensures the existence of $i^\star$ such that $|\Delta_{i^\star}| \leq 2\varepsilon$. Therefore, we have

$$\Delta(\mathsf{Ext}(X,i^\star);U_1) \leq \varepsilon.$$

5 Probabilistic constructions and lower bounds

In this section, we begin by discussing probabilistic constructions of somewhere and probable extractors via the probabilistic method with a uniformly random function. Then, we derive lower bounds on the number of seeds D required by a uniformly random function to be a (k,ε,δ)-probable extractor with non-negligible probability.

As our first main result, we show that, under a very mild constraint on k, ε, and δ, a random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ is a (k,ε,δ)-probable extractor with high probability when $D = C \cdot \frac{n-k}{\varepsilon^2 \cdot \delta}$ for a sufficiently large constant $C > 0$. More precisely, we have the following result.

Theorem 7. *There exist absolute constants $c_1, c_2 > 0$ such that for every $k \leq n-1$ a random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ is a (k,ε,δ)-probable extractor with probability at least 0.99 provided that*

$$k \geq \log\log(2/\delta) + 2\log(1/\varepsilon) + c_1,$$
$$D \geq c_2 \cdot \frac{n-k}{\varepsilon^2 \cdot \delta}.$$

We prove Theorem 7 by relating probable extractors to strong two-source extractors. Then, we invoke the well-known existence result for such objects obtained via the probabilistic method.

Definition 7 ($(n_1, k_1, n_2, k_2, \varepsilon)$-strong two-source extractor)**.** *A function* $\mathsf{Ext} : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}$ *is an* $(n_1, k_1, n_2, k_2, \varepsilon)$-strong two-source extractor *if*

$$\mathop{\mathbb{E}}_{Y}[\Delta(\mathsf{Ext}(X, Y); U_1)] \leq \varepsilon$$

for every independent (n_1, k_1)*-sources* X *and* (n_2, k_2)*-sources* Y*.*

Note that a function Ext is an $(n_1, k_1, n_2, k_2, \varepsilon)$-strong two-source extractor if and only if

$$\mathop{\mathbb{E}}_{i \leftarrow S}[\Delta(\mathsf{Ext}(X, i); U_1)] \leq \varepsilon \tag{21}$$

for every set $S \subseteq \{0,1\}^{n_2}$ of size $|S| = 2^{k_2}$. This observation leads to the following result.

Lemma 7. *Let* $d = \log D$*. Then, every* $(n, k, d, d - \log(1/\delta), \varepsilon)$*-strong two-source extractor* $\mathsf{Ext} : \{0,1\}^n \times [D] \to \{0,1\}$ *is a* (k, ε, δ)*-probable extractor.*

Proof. Suppose Ext is not a (k, ε, δ)-probable extractor. Then, there exists an (n, k)-source X such that

$$\Pr_{i \leftarrow [D]}[\Delta(\mathsf{Ext}(X, i); U_1) > \varepsilon] \geq \delta.$$

This implies that there is a set $S \subseteq [D]$ of size $|S| \geq \delta \cdot D = 2^{d - \log(1/\delta)}$ such that $\Delta(\mathsf{Ext}(X, i); U_1) > \varepsilon$ for every $i \in S$. Taking into account (21), this shows that Ext is not an $(n, k, d, d - \log(1/\delta), \varepsilon)$-strong two-source extractor. □

It is known [DO03, Theorem 1] that, for integers $n_1 > k_1$ and $n_2 > k_2$, a uniformly random function $F : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}$ is an $(n_1, k_1, n_2, k_2, \varepsilon)$-strong two-source extractor with probability at least 0.99 if

$$k_1 \geq \log(n_2 - k_2) + 2\log(1/\varepsilon) + c_1,$$
$$k_2 \geq \log(n_1 - k_1) + 2\log(1/\varepsilon) + c_2,$$

for some absolute constants $c_1, c_2 > 0$. Setting $n_1 = n$, $k_1 = k$, $n_2 = d$, and $k_2 = d - \log(1/\delta)$ and combining the above with Lemma 7 implies the statement of Theorem 7 for all $\delta = 2^{-b}$ for some integer $b \geq 1$ and integers $k \leq n - 1$. The statement can be immediately extended to all $\delta \in (0, 1]$ and, say, all $k \leq n - 1$ (with slightly larger constants c_1 and c_2) by noting that every (k, ε, δ)-probable extractor is also a $(k', \varepsilon, \delta')$-probable extractor for $\delta' > \delta$ and $k' > k$, and that for every δ there is $b \geq 1$ such that $\delta/2 \leq 2^{-b} \leq \delta$.

Remark 2. One might wonder why we obtained a probabilistic construction for (k, ε, δ)-probable extractors via a connection to strong two-source extractors and not by applying the probabilistic method more directly. The reason for this is that, in contrast with strong extractors, the fact that $F : \{0,1\}^n \times [D] \to \{0,1\}$ satisfies

$$\Pr_{i \leftarrow [D]}[\Delta(F(X, i); U_1) > \varepsilon] < \delta$$

for every *flat* (n, k)-source X is *not* sufficient to argue that F is a (k, ε, δ)-probable extractor. This means there is not a natural small set of sources we can restrict our attention to for the union bound.

We restate and prove Theorem 6 below.

Theorem (Theorem 6, restated)**.** *For any $\delta = \delta(n) \in (0,1]$ and large enough n, suppose that $k \le n - 100$, $2^{-0.24(n+k)} \le \varepsilon \le c_0$ for a sufficiently small constant $c_0 > 0$, and*

$$D \le \frac{n-k}{100 \cdot \varepsilon^2 \cdot \delta}\,.$$

Then, a uniformly random function $F : \{0,1\}^n \times [D] \to \{0,1\}$ is not *a (k, ε, δ)-probable extractor with probability at least $1 - 2^{-2^{\Omega(n)}}$.*

We prove the statement above for the particular case of (k,ε)-somewhere extractors (i.e., Theorem 3). The statement for general (k,ε,δ)-probable extractors follows immediately because the function F restricted to the first $\delta D \le \frac{n-k}{100\cdot\varepsilon^2}$ seeds (which is still a uniformly random function) must be a (k,ε)-somewhere extractor in order for F to be a (k,ε,δ)-probable extractor.

We will take X to be the uniform distribution over some set $\mathcal{X}_F$ which depends on the choice of F, and show the following properties hold for large enough n with high probability over the choice of F: (i) we have $|\mathcal{X}_F| \ge K$, so that X is an (n,k)-source, and (ii) it holds that $\Delta(F(X,i);U_1) > \varepsilon$ for every $i \in [D]$. To that end, for $t = D/2 - c\sqrt{(n-k)D}$ where $c = \sqrt{\frac{\ln 2}{36}}$, we define the set

$$\mathcal{X}_F = \{x \in [N] : \|F(x)\|_1 \le t\},$$

and let X be uniformly distributed over $\mathcal{X}_F$. The desired result follows immediately by combining the two lemmas below with a union bound. Throughout the remainder of this section, we set $N = 2^n$, $K = 2^k$, and $g = n - k$. Without loss of generality, we may also assume that $\varepsilon < 1/100$ and $D = \frac{g}{100\cdot\varepsilon^2} \ge 100g$.

Lemma 8. *With probability at least $1 - 2^{-N^{\Omega(1)}}$ over the choice of F, it holds that $|\mathcal{X}_F| \ge \frac{1}{2}\sqrt{KN} \ge K$. Hence, X is an (n,k)-source with probability at least $1 - 2^{-N^{\Omega(1)}}$.*

Proof. Note that $|\mathcal{X}_F|$ is distributed according to a binomial distribution with N trials and success probability $2^{-D} \cdot V_t$. Therefore,

$$\begin{aligned}\mathbb{E}[|\mathcal{X}_F|] &= N \cdot 2^{-D} \cdot V_t \\ &\ge N \cdot \exp\bigl(-9/2 \cdot 4c^2 g\bigr) \\ &= \sqrt{KN},\end{aligned}$$

where the first inequality follows from Lemma 2 with $\gamma = 2c\sqrt{g/D}$ (note that $\gamma \le 2c \le 1/2$ and $\gamma^2 D \ge 6$ since $g \ge 100$) and the second equality holds by the choice of c. By the Chernoff bound[5], it follows that

$$\begin{aligned}\Pr\left[|\mathcal{X}_F| < \frac{1}{2}\cdot\sqrt{KN}\right] &\le \Pr\left[|\mathcal{X}_F| < \frac{1}{2}\cdot\mathbb{E}[|\mathcal{X}_F|]\right] \\ &\le \exp\left(-\frac{\sqrt{KN}}{8}\right) = 2^{-N^{\Omega(1)}}.\end{aligned}$$

□

[5]The version of the Chernoff bound we use here states that $\Pr[X \le (1-\delta)\,\mathbb{E}[X]] \le \exp\bigl(-\frac{\delta^2\,\mathbb{E}[X]}{2}\bigr)$ for $X = \sum_{i=1}^N X_i$ with $X_i \in \{0,1\}$ independent and any $\delta \in [0,1]$.

Lemma 9. *Suppose $\mathcal{X}_F$ has size $|\mathcal{X}_F| \geq \frac{1}{2}\sqrt{KN}$. Then, we have $\Delta(F(X,i);U_1) > \varepsilon$ simultaneously for all $i \in [D]$ with probability at least $1 - 2^{-N^{\Omega(1)}}$ over the choice of F.*

Proof. We prove that this holds with probability at least $1 - 2^{-N^{\Omega(1)}}$ over the choice of F in the case $i = 1$. The reasoning is analogous for all $i \in [D]$, and a union bound over all $i \in [D]$ then yields the desired result, since $D \leq N^{0.99}$ for large enough N.

For $b \in \{0,1\}$, define $\mathcal{X}_{F,b} = \{x \in \mathcal{X}_F : F(x,1) = b\}$. Then, we have

$$\Delta(F(X,1);U_1) = \frac{1}{2} \cdot \left| \frac{|\mathcal{X}_{F,0}| - |\mathcal{X}_{F,1}|}{|\mathcal{X}_F|} \right|. \tag{22}$$

Note that $|\mathcal{X}_{F,0}| - |\mathcal{X}_{F,1}| = \sum_{x \in \mathcal{X}_F} Z_x$, where the Z_x are i.i.d. and $Z_x = 1$ if $x \in \mathcal{X}_{F,0}$ or $Z_x = -1$ if $x \in \mathcal{X}_{F,1}$. Observe also that

$$\Pr[Z_x = 1 | x \in \mathcal{X}_F] = \frac{\sum_{i=0}^{t} \binom{D-1}{i}}{V_t}, \tag{23}$$

$$\Pr[Z_x = -1 | x \in \mathcal{X}_F] = \frac{\sum_{i=0}^{t-1} \binom{D-1}{i}}{V_t}, \tag{24}$$

recalling that V_t denotes the volume of the Hamming ball with radius t. As a result, defining $\binom{a}{b} = 0$ for $b < 0$, we have

$$\begin{aligned}
\mathbb{E}[|\mathcal{X}_{F,0}| - |\mathcal{X}_{F,1}|] &= |\mathcal{X}_F| \cdot \frac{\sum_{i=0}^{t} \binom{D-1}{i} - \sum_{i=0}^{t-1} \binom{D-1}{i}}{V_t} \\
&= |\mathcal{X}_F| \cdot \frac{\sum_{i=0}^{t} \left[\binom{D-1}{i} - \binom{D-1}{i-1} \right]}{V_t} \\
&\geq |\mathcal{X}_F| \cdot \frac{\sum_{i=0}^{t} \left(1 - \frac{2t}{D}\right) \binom{D}{i}}{V_t} \\
&= |\mathcal{X}_F| \left(1 - \frac{2t}{D}\right) \\
&= |\mathcal{X}_F| \cdot 2c\sqrt{g/D}.
\end{aligned}$$

The first equality follows from (23) and (24). The first inequality holds because

$$\binom{D-1}{i} - \binom{D-1}{i-1} = \left(1 - \frac{2i}{D}\right)\binom{D}{i} \geq \left(1 - \frac{2t}{D}\right)\binom{D}{i}$$

for all $0 \leq i \leq t$. The last equality follows from the choice of t.

By Hoeffding's inequality, we have that

$$\begin{aligned}
\Pr\left[|\mathcal{X}_{F,0}| - |\mathcal{X}_{F,1}| \leq |\mathcal{X}_F| \cdot c\sqrt{g/D} \right] &\leq \Pr\left[\frac{|\mathcal{X}_{F,0}| - |\mathcal{X}_{F,1}|}{|\mathcal{X}_F|} \leq \frac{\mathbb{E}[|\mathcal{X}_{F,0}| - |\mathcal{X}_{F,1}|]}{|\mathcal{X}_F|} - c\sqrt{g/D} \right] \\
&\leq \exp\left(-\frac{|\mathcal{X}_F| c^2 g}{2D} \right) = 2^{-N^{\Omega(1)}},
\end{aligned} \tag{25}$$

where the last equality holds because $D = O(g/\varepsilon^2) = O((KN)^{0.49})$ and $|\mathcal{X}_F| \geq \frac{1}{2}\sqrt{KN}$ by hypothesis. From (22) and (25), we conclude that

$$\Delta(F(X,1);U_1) \geq c\sqrt{g/D} > \varepsilon$$

holds with probability at least $1 - 2^{-N^{\Omega(1)}}$. □

Combining Lemmas 8 and 9 with a union bound immediately implies Theorem 3, and therefore also Theorem 6 by the discussion above.

References

[AOR+19] Divesh Aggarwal, Maciej Obremski, João Ribeiro, Luisa Siniscalchi, and Ivan Visconti. How to extract useful randomness from unreliable sources. Cryptology ePrint Archive, Report 2019/1156, 2019. `https://eprint.iacr.org/2019/1156`.

[BCD+18] Avraham Ben-Aroya, Eshan Chattopadhyay, Dean Doron, Xin Li, and Amnon Ta-Shma. A new approach for constructing low-error, two-source extractors. In *Proceedings of the 33rd Computational Complexity Conference*, CCC '18, pages 3:1–3:19, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[BKS+10] Boaz Barak, Guy Kindler, Ronen Shaltiel, Benny Sudakov, and Avi Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. *J. ACM*, 57(4):20:1–20:52, May 2010.

[BRSW12] Boaz Barak, Anup Rao, Ronen Shaltiel, and Avi Wigderson. 2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction. *Annals of Mathematics*, pages 1483–1543, 2012.

[CGL16] Eshan Chattopadhyay, Vipul Goyal, and Xin Li. Non-malleable extractors and codes, with their many tampered extensions. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 285–298, New York, NY, USA, 2016. ACM.

[Coh15] Gil Cohen. Local correlation breakers and applications to three-source extractors and mergers. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 845–862, Oct 2015.

[CZ19] Eshan Chattopadhyay and David Zuckerman. Explicit two-source extractors and resilient functions. *Annals of Mathematics*, 189(3):653–705, 2019.

[DKSS13] Zeev Dvir, Swastik Kopparty, Shubhangi Saraf, and Madhu Sudan. Extensions to the method of multiplicities, with applications to Kakeya sets and mergers. *SIAM Journal on Computing*, 42(6):2305–2328, 2013.

[DO03] Yevgeniy Dodis and Roberto Oliveira. On extracting private randomness over a public channel. In Sanjeev Arora, Klaus Jansen, José D. P. Rolim, and Amit Sahai, editors, *Approximation, Randomization, and Combinatorial Optimization.. Algorithms and Techniques*, pages 252–263, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[DR08] Zeev Dvir and Ran Raz. Analyzing linear mergers. *Random Structures & Algorithms*, 32(3):334–345, 2008.

[DS07] Zeev Dvir and Amir Shpilka. An improved analysis of linear mergers. *computational complexity*, 16(1):34–59, May 2007.

[DW11] Zeev Dvir and Avi Wigderson. Kakeya sets, new mergers, and old extractors. *SIAM Journal on Computing*, 40(3):778–792, 2011.

[KY15] Philip Klein and Neal E. Young. On the number of iterations for Dantzig–Wolfe optimization and packing-covering approximation algorithms. *SIAM Journal on Computing*, 44(4):1154–1172, 2015.

[Li11] Xin Li. Improved constructions of three source extractors. In *2011 IEEE 26th Annual Conference on Computational Complexity*, pages 126–136, June 2011.

[Li13a] Xin Li. Extractors for a constant number of independent sources with polylogarithmic min-entropy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 100–109, Oct 2013.

[Li13b] Xin Li. New independent source extractors with exponential improvement. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 783–792, New York, NY, USA, June 2013. ACM.

[Li15] Xin Li. Three-source extractors for polylogarithmic min-entropy. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 863–882, Oct 2015.

[Li16] Xin Li. Improved two-source extractors, and affine extractors for polylogarithmic entropy. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 168–177, Oct 2016.

[LRVW03] Chi-Jen Lu, Omer Reingold, Salil Vadhan, and Avi Wigderson. Extractors: Optimal up to constant factors. In *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, STOC '03, pages 602–611, New York, NY, USA, 2003. ACM.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *Journal of Computer and System Sciences*, 52(1):43 – 52, 1996.

[Rao07] Anup Rao. An exposition of Bourgain's 2-source extractor. 2007.

[Rao09] Anup Rao. Extractors for a constant number of polynomially small min-entropy independent sources. *SIAM Journal on Computing*, 39(1):168–194, 2009.

[Raz05] Ran Raz. Extractors with weak random seeds. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 11–20, New York, NY, USA, 2005. ACM.

[RT00] Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24, 2000.

[Ta-96] Amnon Ta-Shma. On extracting randomness from weak random sources (extended abstract). In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 276–285, New York, NY, USA, 1996. ACM.

[Zuc06] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pages 681–690, New York, NY, USA, 2006. ACM.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il