SIAM J. COMPUT.
Vol. 50, No. 3, pp. STOC16-30–STOC16-67

TWO-SOURCE DISPERSERS FOR POLYLOGARITHMIC ENTROPY AND IMPROVED RAMSEY GRAPHS*

GIL COHEN†

Abstract. In his 1947 paper that inaugurated the probabilistic method, Erdős proved the existence of $(2+o(1))\log n$-Ramsey graphs on n vertices. Matching Erdős's result with a constructive proof is considered a central problem in combinatorics and has gained significant attention in the literature. The state-of-the-art result was obtained in the celebrated paper by Barak et al. [*Ann. of Math.* (2), 176 (2012), pp. 1483–1543], who constructed a $2^{2^{(\log\log n)^{1-\alpha}}}$-Ramsey graph for some universal constant $\alpha > 0$. In this work, we significantly improve the result of Barak et al. and construct $2^{(\log\log n)^c}$-Ramsey graphs, for some universal constant c. In the language of theoretical computer science, this resolves the problem of explicitly constructing dispersers for two n-bit sources with entropy polylog(n). In fact, our disperser is a zero-error disperser that outputs a constant fraction of the entropy. Previously, such dispersers could only support entropy $\Omega(n)$.

Key words. Ramsey graphs, two-source dispersers, explicit constructions

AMS subject classifications. 05D10, 68Q87

DOI. 10.1137/16M1096219

1. Introduction.

Ramsey theory is a branch of combinatorics that studies the unavoidable presence of local structure in globally unstructured objects. In the paper that pioneered this field of study, Ramsey [35] considered an instantiation of this phenomena in graph theory.

DEFINITION 1.1 (Ramsey graphs). *A graph on n vertices is called k-Ramsey if it contains no clique or independent set of size k.*

Ramsey showed that there does not exist a graph on n vertices that is $\log(n)/2$-Ramsey. In his influential paper that inaugurated the probabilistic method, Erdős [18] complemented Ramsey's result and showed that most graphs on n vertices are $(2 + o(1))\log n$-Ramsey. Unfortunately, Erdős's argument is nonconstructive, and one does not obtain from Erdős's proof an example of a Ramsey graph with such parameters. Erdős offered a $100 prize for matching his result, up to any multiplicative constant factor, by a constructive proof, that is, coming up with an explicit construction of an $O(\log n)$-Ramsey graph. Erdős's challenge gained significant attention in the literature. In Table 1 we give a summary of known explicit constructions. Other works studied the difficulty of constructing Ramsey graphs [23] and suggested routes toward constructing improved Ramsey graphs [24].

The notion of explicitness was formalized in the computational era. While, classically, a succinct mathematical formula was widely considered to be an explicit description, complexity theory suggests a more relaxed, and arguably more natural, interpretation of explicitness. An object is deemed explicit if one can efficiently con-

*Received by the editors September 29, 2016; accepted for publication (in revised form) October 2, 2017; published electronically October 21, 2019. An extended abstract of this work was published in the Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS). https://doi.org/10.1137/16M1096219

Funding: This work was partially supported by an ISF grant and by the I-CORE program of the planning and budgeting committee.

†Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Current address: Department of Computer Science, Princeton University, Princeton, NJ 08540 (gilc@princeton.edu).

struct that object from scratch. More specifically, a graph on n vertices is explicit if given the labels of any two vertices u, v, one can efficiently determine whether there is an edge connecting u, v in the graph. Since the description of u, v consists of $2 \log n$ bits, quantitatively, for efficiency we require that the running-time for determining connectivity between any two vertices be polylog(n).

Ramsey graphs have an analogous definition for bipartite graphs. A bipartite graph on two sets of n vertices is bipartite k-Ramsey if it has no $k \times k$ complete or empty bipartite subgraph. One can show that a bipartite Ramsey graph induces a Ramsey graph with comparable parameters. Thus, constructing bipartite Ramsey graphs is at least as hard as constructing Ramsey graphs, and it was believed to be a strictly harder problem. Nevertheless, the best known construction of Ramsey graphs is in fact bipartite. Furthermore, Erdős's argument holds as is for bipartite graphs.

Building on [5], in their celebrated paper, Barak et al. [6] demonstrated an explicit bipartite $k(n)$-Ramsey graph on n vertices with $k(n) = 2^{2^{(\log \log n)^{1-\alpha}}}$, where $\alpha > 0$ is some small universal constant. In particular, $k(n) = 2^{o(\log n)}$ is subexponential in the desired value, namely, $\log n$. In this paper we give an explicit construction of a bipartite $k(n)$-Ramsey graph with $k(n)$ that is quasi-polynomial in the desired value.[1]

TABLE 1
Summary of Ramsey graph constructions from the literature.

Construction	$k(n)$	Bipartite
[18] (nonconstructive)	$2 \log n$	✓
[1]	$n^{\log_5 2}$	
[32]	$n^{1/3}$	
[20]	$n^{o(1)}$	
[13]	$2^{O((\log n)^{3/4} \cdot (\log \log n)^{1/4})}$	
[21, 33, 2, 25, 3]	$2^{O(\sqrt{\log n \cdot \log \log n})}$	
The Hadamard matrix (folklore)	$\sqrt{n}$	✓
[34]	$n^{1/2-o(1)}$	✓
[5]	$n^{o(1)}$	✓
[6]	$2^{2^{(\log \log n)^{1-\alpha}}}$	✓
This work	$2^{(\log \log n)^{O(1)}}$	✓

THEOREM 1.2 (Ramsey graphs). *There exists an explicit bipartite* $2^{(\log \log n)^c}$*-Ramsey graph on* n *vertices, where* c *is some universal constant.*

We remark that the constant c in Theorem 1.2 as well as the constant in the exponent of the polylog(n) running-time are not too large, though we made no attempt at bounding them. On the other hand, the algorithm that generates our Ramsey graph is fairly involved and does not have a short and simple description. Presenting the algorithm, even without taking into account the (highly involved) building blocks from the literature that we use, will require some preparation.

[1] A function $f\colon \mathbb{N} \to \mathbb{N}$ is *quasi-polynomial* if there exist constants c, m_0 such that $f(m) \leq 2^{(\log m)^c}$ for all $m > m_0$.

It is worth mentioning that the graph that we construct has a stronger property than being Ramsey. Namely, for $k = 2^{(\log\log n)^c}$, any k by k bipartite subgraph has a relatively large bipartite subgraph of its own with edge density close to $1/2$. An analogous property holds even in the multicolored setting. This stronger property is related to the well-known two-source extractors problem from theoretical computer science. In what follows we move to present our results in the language of computer science. We do so mainly because the techniques we apply are most naturally presented from that perspective.

1.1. Two-source zero-error dispersers, extractors, and subextractors. In the language of theoretical computer science, Theorem 1.2 translates to a disperser for two independent n-bit sources with entropy $O(\log^c n)$. We first recall some basic definitions.

DEFINITION 1.3 (statistical distance). *The* statistical distance *between two distributions X, Y on a common domain D is defined by*

$$\mathsf{SD}(X,Y) = \max_{A \subseteq D} \{ |\, \mathbf{Pr}[X \in A] - \mathbf{Pr}[Y \in A]\, | \}.$$

If $\mathsf{SD}(X,Y) \le \varepsilon$ we say that X is ε-close to Y (and, of course, Y is ε-close to X) and write $X \sim_\varepsilon Y$.

DEFINITION 1.4 (min-entropy). *The* min-entropy *of a random variable X is defined by*

$$\mathbf{H}_\infty(X) = \min_{x \in \mathsf{supp}(X)} \log_2 \left(\frac{1}{\mathbf{Pr}[X = x]} \right).$$

If X is supported on $\{0,1\}^n$, we define the min-entropy rate *of X by $\mathbf{H}_\infty(X)/n$. In such a case, if X has min-entropy k or more, we say that X is an (n,k)-weak-source or simply an (n,k)-source. We sometimes abbreviate and simply say* entropy *(resp.,* entropy rate*) instead of min-entropy (resp., min-entropy rate). This should cause no confusion as the only measure of entropy used in this paper is min-entropy.*

DEFINITION 1.5 (two-source zero-error dispersers). *A function $\mathsf{Disp}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ is called a* two-source zero-error disperser *for entropy k if for any two independent (n,k)-sources X, Y, it holds that*

$$\mathsf{supp}(\mathsf{Disp}(X,Y)) = \{0,1\}^m.$$

Note that a two-source zero-error disperser for entropy k, with a single output bit (namely, $m = 1$), is equivalent to a bipartite 2^k-Ramsey graph on 2^n vertices on each side. Constructing two-source dispersers for polylogarithmic entropy is considered a central problem in pseudorandomness that we resolve in this paper. Indeed, a bipartite $2^{\mathrm{poly}(\log\log n)}$-Ramsey graph on n vertices is equivalent to a disperser for entropy $\mathrm{polylog}(n)$. From the point of view of dispersers, it is easier to see how challenging Erdős's goal of constructing $O(\log n)$-Ramsey graphs is. Indeed, these are equivalent to dispersers for entropy $\log_2(n) + O(1)$. Even a disperser for entropy $O(\log n)$ does not quite meet Erdős's goal as it translates to a $\mathrm{polylog}(n)$-Ramsey graph.

While Theorem 1.2 already yields a two-source zero-error disperser for polylogarithmic entropy, it is desired to construct dispersers with many output bits. Our construction has this property.

THEOREM 1.6 (two-source zero-error dispersers). *There exists an explicit two-source zero-error disperser for n-bit sources having entropy $k = \text{polylog}(n)$, with $m = k^{\Omega(1)}$ output bits.*

Theorem 1.6 gives an explicit zero-error disperser for polylogarithmic entropy with many output bits. Prior to this work, the state-of-the-art zero-error disperser with a super constant number of output bits, due to Gabizon and Shaltiel [22], required entropy $k = \Omega(n)$. In fact, partially motivated by applications to data structures [19], in [22] a stronger variant of a two-source zero-error disperser was constructed, in which every element in the range is obtained with probability at least $\delta = \delta(n)$. Our construction has this property as well. In fact, a stronger property holds. Before discussing this stronger property, we remark that by applying a result of [22], one can increase the output length of the disperser from Theorem 1.6 to $m = \Omega(k)$ without asymptotic loss of parameters.

As previously mentioned, our construction has a stronger property than merely being a disperser. To present this property, we start by recalling the notion of a two-source extractor, introduced by Chor and Goldreich [12].

DEFINITION 1.7 (two-source extractors). *A function* $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is called a* two-source extractor *for entropy k, with error guarantee ε, if for any two independent (n,k)-sources X, Y, it holds that $\mathsf{Ext}(X,Y)$ is ε-close to uniform.*

Chor and Goldreich [12] proved that there exist two-source extractors with error guarantee ε for entropy $k = \log(n) + 2\log(1/\varepsilon) + O(1)$ with $m = 2k - 2\log(1/\varepsilon) - O(1)$ output bits. A central open problem in pseudorandomness is to match this existential proof with an explicit construction having comparable parameters. Unfortunately, even after almost 30 years, little progress has been made even when setting $m = 1$.

In their paper, Chor and Goldreich gave an explicit construction of a two-source extractor for entropy $(0.5 + \delta)n$, where $\delta > 0$ is any fixed constant. Although this is very far from what is obtained by the existential argument, it took almost 20 years before any improvement was made. Bourgain [9] constructed a two-source extractor for entropies $(1/2 - \alpha) \cdot n$, where $\alpha > 0$ is some small universal constant. An incomparable result was obtained by Raz [38], who required one source to have entropy $(0.5 + \delta)n$, where $\delta > 0$ is any fixed constant, but allowed the other source to have entropy $O(\log n)$. Several weaker variants of two-source extractors [26, 36] and conditional two-source extractors [12, 8] were constructed in the literature, but even these constructions only support linear entropy.

In this paper we construct a pseudorandom object that is stronger than a two-source zero-error disperser yet is weaker than a two-source extractor. Informally speaking, this is a function with the following property. "In" any two independent weak-sources there exist two independent weak-sources with a comparable amount of entropy to the original sources, restricted to which, the function is close to uniform. To give a formal definition, we first recall the definition of a subsource, introduced in [5].

DEFINITION 1.8 (subsources). *Given random variables X and X' on $\{0,1\}^n$, we say that X' is a subsource of X and write $X' \subset X$ if there exists a set $A \subseteq \{0,1\}^n$ such that $X' = X \mid \{X \in A\}$. That is, for every $a \in A$, $\mathbf{Pr}[X' = a]$ is defined by $\mathbf{Pr}[X = a \mid X \in A]$ and for $a \notin A$, $\mathbf{Pr}[X' = a] = 0$.*

DEFINITION 1.9 (two-source subextractors). *A function* $\mathsf{SubExt}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is called a* two-source subextractor *for outer-entropy k_{out} and inner-entropy k_{in}, with error guarantee ε, if the following holds. For any independent (n, k_{out})-*

sources X, Y, there exist min-entropy k_{in} subsources $X' \subset X$, $Y' \subset Y$, such that $\mathsf{SubExt}(X', Y')$ *is ε-close to uniform.*

Although we are not aware of the definition of two-source subextractors made explicit in previous works, we note that the two-source disperser constructed by Barak et al. [5] is in fact a two-source subextractor. More precisely, for any constant $\delta > 0$, the authors construct a two-source subextractor for outer-entropy δn and inner-entropy $\mathrm{poly}(\delta)n$. On the other hand, by a careful inspection, the state-of-the-art two-source disperser by Barak et al. [6] does not seem to be a subextractor.

The main theorem proved in this paper is the following.

THEOREM 1.10 (two-source subextractors). *There exist a universal constant c and an explicit n-bit two-source subextractor for outer-entropy $k_{\mathsf{out}} = \log^c n$ and inner-entropy $k_{\mathsf{in}} = k_{\mathsf{out}}^{\Omega(1)}$, with $m = k_{\mathsf{in}}^{\Omega(1)}$ output bits and error guarantee $\varepsilon = 2^{-k_{\mathsf{in}}^{\Omega(1)}}$.*

We note that a subextractor for outer-entropy k_{out} with m output bits and error guarantee ε is a zero-error disperser for entropy k_{out} with $\min(m, \log(1/\varepsilon)) - 1$ output bits (the dependence in the inner-entropy k_{in} follows implicitly due to the general fact that $m \le k_{\mathsf{in}}$). Indeed, one can simply truncate the output of the subextractor to be short enough so that the error will be sufficiently small to guarantee that any possible output is obtained. In particular, a subextractor for outer-entropy k_{out} and inner-entropy $k_{\mathsf{in}} = 1$, with error guarantee $\varepsilon < 1/2$, induces a bipartite $2^{k_{\mathsf{out}}}$-Ramsey graph. Thus, Theorem 1.10 readily implies Theorems 1.2 and 1.6.

1.2. Subsequent work. In an exciting subsequent work, Chattopadhyay and Zuckerman [11] gave a construction of a two-source extractor for polylog(n)-entropy based on a very different set of ideas than ours. The error of their extractor is polynomially small in n and the number of output bits is 1. The latter was improved soon after by Li [29] using similar techniques, with the same error parameter. As extractors with one output bit yield Ramsey graphs, the work of [11] gives a second and very different construction of Ramsey graphs matching our parameters.

Following [11], mainly for applications to Ramsey graphs, a line of research was devoted to reducing the entropy required by a two-source extractor from polylog(n) to $O(\log n)$. In [17] it was noted that existing techniques will not yield extractors with entropy lower than $\log^2 n$. The authors introduced the notion of an *independence-preserving merger* and used their construction of such a pseudorandom object to devise an extractor for $O(1/\delta)$ sources, each with min-entropy $(\log n)^{1+\delta}$. In a follow-up work Chattopadhyay and Li [10] gave an improved construction of (a variant of) an independence-preserving merger and used that to obtain an extractor for $O(1)$ sources each with entropy $\log n \cdot 2^{\sqrt{\log \log n}} = (\log n)^{1+o(1)}$. In an independent work [15], a five-source extractor for entropy $\log n \cdot 2^{\sqrt{\log \log n}}$ was constructed. Ben-Aroya, Doron, and Ta-Shma [7] then got the number of sources down to two, with similar min-entropy. In fact, their result can be viewed as a lossless reduction from two-source extractors to nonmalleable extractors, improving upon a lossy reduction that was introduced in [11] and on lossless reduction from [15] that requires five rather than two sources. In subsequent works [16, 31], improved nonmalleable extractors were constructed, and by appealing to the Ben-Aroya et al. reduction, these results yield an explicit bipartite k-Ramsey graph for $k = (\log n)^{O(\log \log \log n)}$ [31].

1.3. Organization of this paper. In section 2 we give an informal overview of the challenge-response mechanism. Section 3 contains a comprehensive and detailed overview of our construction and analysis. These two sections are meant only for

building intuition. The reader may freely skip these sections at any point as we make no use of the results that appear in them.

In section 4 we give some preliminary definitions and results that we need. Section 5 contains the formal description of the challenge-response mechanism. In section 6 we present the notions of entropy-trees and tree-structured sources. Finally, in section 7 we give the formal construction of our subextractor, and we analyze it in section 8.

2. Overview of the challenge-response mechanism. Our subextractor construction is based on the challenge-response mechanism that was introduced in [5] and refined in [6]. As we are aiming for a self-contained paper, in this section we explain how this powerful mechanism works. Further, our presentation is somewhat more abstract than [5, 6], which we believe may contribute to the clarity of the exposition. To illustrate the way the mechanism works, we give a toy example in section 2.4.

Before presenting the challenge-response mechanism, we give the definition for the deficiency of a subsource [5]. Let X' be a subsource of X, and let A be the set such that $X' = X \mid \{X \in A\}$. Then, we say that X' is a deficiency d subsource of X if $\mathbf{Pr}[X \in A] \geq 2^{-d}$.

2.1. Motivating the challenge-response mechanism. We start by recalling the notation of a block-source [12].

DEFINITION 2.1. *Let n be an even integer. A random variable X on n-bit strings is called an* (n,k)-block-source*, or simply a k-block-source, if the following hold:*

- $\mathbf{H}_\infty(\mathsf{left}(X)) \geq k$*, where* $\mathsf{left}(X)$ *is the length $n/2$ prefix of X.*
- *For any $x \in \mathsf{supp}(\mathsf{left}(X))$ it holds that $\mathbf{H}_\infty(\mathsf{right}(X) \mid \{\mathsf{left}(X) = x\}) \geq k$, where* $\mathsf{right}(X)$ *is the length $n/2$ suffix of X.*

Following a long line of research [12, 4, 38, 5, 37, 27, 28], in a recent breakthrough, Li [30] gave a construction of an extractor BExt for two n-bit sources, where the first source is a polylog(n)-block-source and the second is a weak-source with min-entropy polylog(n) (see Theorem 4.1). In particular, Li obtained a three-source extractor for polylogarithmic entropy, but his result is stronger than that, and we are using this stronger property. As our goal is to construct a two-source subextractor for outer-entropy polylog(n), a first attempt would be to show that any source X with entropy polylog(n) has a subsource X' that is a polylog(n)-block-source. If this assertion were to be true, then BExt would have been a two-source subextractor.

This, however, is clearly not the case. Consider, for example, a source X that has all of its entropy concentrated in its right-block $\mathsf{right}(X)$. Namely, $\mathsf{left}(X)$ is fixed to some constant and $\mathsf{right}(X)$ has min-entropy k. Clearly, $\mathbf{H}_\infty(X) \geq k$, yet no subsource of X is even a 1-block-source.

Such an example can only hold when the entropy is no larger than $n/2$. Indeed, informally speaking, one cannot squeeze, say, $0.6n$ entropy to the $n/2$ bits of $\mathsf{right}(X)$. Restricting ourselves, for the moment, to the very high entropy regime, we ask whether this example is the only problematic example. In particular, is it true that any n-bit source with min-entropy $0.6n$ is a block-source? The answer to this question is still no. Nevertheless, one can show that any $(n, 0.6n)$-weak-source has a low-deficiency subsource that is a $0.1n$-block-source. This simple observation will be important for us later on.

Going back to the example above, if only there were a magical algorithm that, given a single sample $x \sim X$, would have been able to determine correctly whether or not $\mathsf{left}(X)$ is fixed to a constant, then we would have been in better shape as

we would have known to "concentrate our efforts" on $\mathsf{right}(X)$. Such an algorithm, however, is too much to hope for. Indeed, given a single sample $x \sim X$, one simply cannot tell whether the left block of X is fixed or not. Still, the powerful challenge-response mechanism allows one to almost accomplish this task using an additional independent sample. In the next section we present a slightly informal version of the challenge-response mechanism. A formal treatment of the actual mechanism is given in section 5.

2.2. The challenge-response mechanism. We start this section by presenting a dream version of the challenge-response mechanism.

The challenge-response mechanism—dream version. For integers $\ell < n$, a dream version of the *challenge-response mechanism* would be a $\mathrm{poly}(n)$-time computable function

$$\mathsf{DreamResp}\colon \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^\ell \to \{\mathsf{fixed}, \mathsf{hasEntropy}\}$$

with the following property. For any two independent $(n, \mathrm{polylog}(n))$-sources X, Y, and for any function $\mathsf{Challenge}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^\ell$, the following hold:

- If $\mathsf{Challenge}(X,Y)$ is fixed to a constant, then

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X,Y)}[\mathsf{DreamResp}\,(x, y, \mathsf{Challenge}(x,y)) = \mathsf{fixed}] = 1.$$

- If $\mathbf{H}_\infty(\mathsf{Challenge}(X,Y))$ is sufficiently large, then

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X,Y)}[\mathsf{DreamResp}\,(x, y, \mathsf{Challenge}(x,y)) = \mathsf{hasEntropy}] = 1.$$

Note that for any function $\mathsf{Challenge}$, the function $\mathsf{DreamResp}$ distinguishes between the case that $\mathsf{Challenge}(X,Y)$ is fixed and the case that $\mathsf{Challenge}(X,Y)$ has enough entropy. Unfortunately, $\mathsf{DreamResp}$ will remain a dream. The actual challenge-response mechanism requires more from the inputs and has a weaker guarantee on the output. The difference between the dream version and the actual mechanism contributes to why our subextractor is defined the way it is, and so in this section we present the actual challenge-response mechanism (though in a slightly informal manner).

The actual challenge-response mechanism. For integers $\ell < n$, the challenge-response mechanism is a $\mathrm{poly}(n)$-time computable function

$$\mathsf{Resp}\colon \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^\ell \to \{\mathsf{fixed}, \mathsf{hasEntropy}\}$$

with the following property. For any two independent $(n, \mathrm{polylog}(n))$-sources X, Y, and for any function $\mathsf{Challenge}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^\ell$, the following hold:

- If $\mathsf{Challenge}(X,Y)$ is fixed to a constant, then there exist deficiency ℓ subsources $X' \subset X$, $Y' \subset Y$, such that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X',Y')}[\mathsf{Resp}(x, y, \mathsf{Challenge}(x,y)) = \mathsf{fixed}] = 1.$$

- If for any deficiency ℓ subsources $\hat{X} \subset X$, $\hat{Y} \subset Y$ it holds that $\mathbf{H}_\infty(\mathsf{Challenge}(\hat{X}, \hat{Y})) \geq k$, then

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X,Y)}[\mathsf{Resp}(x, y, \mathsf{Challenge}(x,y)) = \mathsf{hasEntropy}] \geq 1 - 2^{-k}.$$

We emphasize the differences between the dream version and the actual challenge-response mechanism. First, even if $\mathsf{Challenge}(X,Y)$ is fixed to a constant, it is not guaranteed that Resp will correctly identify this on any sample from (X,Y). In fact, it is not even guaranteed that Resp will identify this correctly with high probability over the sample. The actual guarantee is that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$, such that given any sample $(x,y) \sim (X',Y')$, Resp will correctly output fixed. As our goal is to construct a subextractor, this is good enough for us, as we can imagine that we are given samples from X', Y' rather than from X, Y for the rest of the analysis (we do have to be careful when dealing with error terms when moving to subsources, but we will ignore this issue for now).

The second thing to notice is that for the challenge-response mechanism to identify the fact that $\mathsf{Challenge}(X,Y)$ has entropy, a stronger assumption is made. Namely, it is not enough that $\mathsf{Challenge}(X,Y)$ has a sufficient amount of entropy, but rather we need that $\mathsf{Challenge}(\hat{X},\hat{Y})$ has enough entropy for *all* low-deficiency subsources $\hat{X} \subset X$, $\hat{Y} \subset Y$. So, informally speaking, for the challenge-response mechanism to "sense" entropy, this entropy must be "robust" in the sense that the entropy exists even when applying $\mathsf{Challenge}$ on all low-deficiency subsources of X, Y. Further, note that, unlike the first case, in the second case Resp introduces a small error.

2.3. The three-types lemma. The challenge-response mechanism is very impressive. However, the mechanism only distinguishes between two extreme cases—no entropy versus high entropy. It is much more desired to be able to distinguish between low entropy versus high entropy. Indeed, altering the example from section 2.1 a bit, what if the entropy in the left block of a source is too low to work with, yet the block is not fixed to a constant, and so the challenge-response mechanism is inapplicable?

The next lemma shows that if we are willing to work with subsources (and we are), then this is a nonissue. Namely, every source has a low-deficiency subsource with a structure suitable for the challenge-response mechanism. We present here a slightly informal version of this lemma. The reader is referred to Lemma 6.7 for the formal statement.

LEMMA 2.2 (the three-types lemma). *For any (n,k)-source X and integer $b < k/2$, there exists a deficiency $b+2$ subsource $X' \subset X$ such that (at least) one of the following holds:*

- *X' is a b-block-source.*
- *$\mathbf{H}_\infty(\mathsf{left}(X')) \geq k - b$.*
- *$\mathsf{left}(X')$ is fixed to a constant, and $\mathbf{H}_\infty(\mathsf{right}(X')) \geq k - b$.*

One should think of $b = o(k)$ that is still "large enough," for example, $b = \sqrt{k}$. In such case, Lemma 2.2, which is a variant of the two-types lemma by Barak et al. [6], states that any source X has a deficiency $\sim \sqrt{k}$ subsource X' with a useful structure. If X' is not a $\sqrt{k}$-block-source, then either essentially all of the entropy already appears in $\mathsf{left}(X')$, or otherwise $\mathsf{left}(X')$ is fixed to a constant and $\mathsf{right}(X')$ has almost all the entropy of X.

As a corollary of the three-types lemma, we conclude the informal assertion made in section 2.1. Namely, if X is an (n,k)-source with $k > n/2$, say $k = (1/2+\alpha)n$ for some constant $\alpha > 0$, then X has a deficiency $\alpha n + O(1)$ subsource that is an $(\alpha n - O(1))$-block-source. Indeed, by applying Lemma 2.2 to X with $b = \alpha n - 1$ we see that the second and third cases of the lemma cannot hold since by this choice of b, $k - b > n/2$ (and the entropy of a random variable on $n/2$ bits cannot exceed its length). Thus, the first case must hold; namely, there is a deficiency $\alpha n + O(1)$

subsource of X that is $(\alpha n - O(1))$-block-source. For ease of readability, in this informal section we typically ignore additive constant loss in deficiency and entropy.

2.4. Playing with the challenge-response mechanism. Lemma 2.2 is an important supplement to the challenge-response mechanism. However, it is still not even clear how the two together can be used to break the "0.5 barrier" discussed in section 2.1—for example, how they together can be used to give a subextractor for outer-entropies $0.4n, \text{polylog}(n)$.

Let us try to see what can be said. Say X is an $(n, 0.4n)$-source. By Lemma 2.2, applied with $b = 0.1n$, there exists a deficiency $0.1n$ subsource X' of X, such that one of the following holds:

- X' is a $0.1n$-block-source.
- $\mathbf{H}_\infty(\mathsf{left}(X')) \geq 0.3n$.
- $\mathsf{left}(X')$ is fixed to a constant, and $\mathbf{H}_\infty(\mathsf{right}(X')) \geq 0.3n$.

Note that in the second case, $\mathsf{left}(X')$ has entropy rate 0.6. Thus, $\mathsf{left}(X')$ has a subsource that is an $\Omega(n)$-block-source. Similarly, in the third case, $\mathsf{right}(X')$ has a subsource that is an $\Omega(n)$-block-source. Thus, we conclude that any $(n, 0.4n)$-source has a subsource $X'' \subset X$ such that at least one of X'', $\mathsf{left}(X'')$, $\mathsf{right}(X'')$ is an $\Omega(n)$-block-source. Further, in the last case, $\mathsf{left}(X'')$ is fixed to a constant.

By the above discussion, even without resorting to the challenge-response mechanism, we know that at least one of $\mathsf{BExt}(X'', Y), \mathsf{BExt}(\mathsf{left}(X''), Y), \mathsf{BExt}(\mathsf{right}(X''), Y)$ is close to uniform. The challenge-response mechanism allows us to obtain something stronger. Although we will not be able to get a subextractor for outer-entropies $0.4n, \text{polylog}(n)$ this way, it is instructive to see the technique being used.

Set BExt to output $\ell = o(k)$ bits, where $k = \text{polylog}(n)$ is the outer-entropy of the second source. Consider the following algorithm.

A toy algorithm. On input $x, y \in \{0,1\}^n$

- Compute $z(x, y) = \mathsf{Resp}(x, y, \mathsf{BExt}(\mathsf{left}(x), y))$.
- If $z = \mathsf{fixed}$ declare that $\mathsf{BExt}(\mathsf{right}(x), y)$ is close to uniform.
- Otherwise, declare that one of $\mathsf{BExt}(x, y)$, $\mathsf{BExt}(\mathsf{left}(x), y)$ is close to uniform.

The above algorithm does not look very impressive. Essentially, it only cuts down our lack of knowledge a bit. That is, instead of declaring that one of three strings is close to uniform, it is able to declare that one of at most two strings is close to uniform. Nevertheless, as mentioned above, it is instructive to see the proof technique applied to this simple toy example. Moreover, as we will see in section 3.2, this algorithm is a special case of an algorithm by [6] that will be important for our construction as well. We now prove that the algorithm's declaration is correct. More precisely, we prove the following.

CLAIM 2.3. *Let X be an $(n, 0.4n)$-source, and let Y be an independent (n, k)-source, with $k = \text{polylog}(n)$. Then, there exist subsources $X' \subset X$, $Y' \subset Y$, such that with probability 1 over $(x, y) \sim (X', Y')$ the declaration of the algorithm is correct.*

The proof of Claim 2.3 showcases the following three facts about low-deficiency subsources. None of these facts is very surprising, but we make extensive use of them throughout the paper, and it is beneficial to see these facts in action on a simple example. Here we give slightly informal (and inaccurate) statements. For the formal statements see Fact 4.3, Fact 4.4, and Lemma 4.6.

FACT 2.4. *Let X be a random variable with min-entropy k, and let X' be a deficiency d subsource of X. Then, $\mathbf{H}_\infty(X') \geq k - d$.*

FACT 2.5. *Let X be a random variable on n-bit strings. Let $f\colon \{0,1\}^n \to \{0,1\}^\ell$ be an arbitrary function. Then, there exist $c \in \{0,1\}^\ell$ and a deficiency ℓ subsource X' of X such that $f(x) = c$ for every $x \in \mathsf{supp}(X')$.*

LEMMA 2.6. *Let X be a k-block-source, and let X' be a deficiency d subsource of X. Then, X' is a $k-d$ block-source.*

Proof of Claim 2.3. We start by applying the three-types lemma as discussed above so as to obtain a subsource $X' \subset X$ with the properties listed above. Namely, at least one of X', $\mathsf{left}(X')$, $\mathsf{right}(X')$ is an $\Omega(n)$-block-source. Further, in the last case, $\mathsf{left}(X')$ is fixed to a constant.

Consider first the case where $\mathsf{left}(X')$ is fixed and $\mathsf{right}(X')$ is an $\Omega(n)$-block-source. Note that in this case, $\mathsf{BExt}(\mathsf{left}(X'), Y)$ is a deterministic function of Y. Since the output length of BExt is ℓ, Fact 2.5 implies that there exists a deficiency ℓ subsource $Y' \subset Y$ such that $\mathsf{BExt}(\mathsf{left}(X'), Y')$ is fixed to a constant. We are now in a position to apply the challenge-response mechanism to conclude that there exist deficiency ℓ subsources $X'' \subset X'$, $Y'' \subset Y'$ such that

$$\mathbf{Pr}\left[z(X'', Y'') = \mathsf{fixed}\right] = 1. \tag{2.1}$$

Recall that $\mathsf{right}(X')$ is an $\Omega(n)$-block-source. Hence, by Lemma 2.6, since X'' is a deficiency $\ell = o(n)$ subsource of X', we have that $\mathsf{right}(X'')$ is also an $\Omega(n)$-block-source. Similarly, since $\mathbf{H}_\infty(Y) = k = \omega(\ell)$, Fact 2.4 implies that $\mathbf{H}_\infty(Y'') = k - 2\ell = \mathrm{polylog}(n)$. Thus, $\mathsf{BExt}(\mathsf{right}(X''), Y'')$ is close to uniform. To summarize, in the case that $\mathsf{left}(X')$ is fixed, there exist subsources $X'' \subset X$, $Y'' \subset Y$ on which the algorithm correctly declares that $\mathsf{BExt}(\mathsf{right}(X''), Y'')$ is close to being uniformly distributed.

Consider now the case where $\mathsf{left}(X')$ is not fixed. Thus, at least one of X', $\mathsf{right}(X')$ is an $\Omega(n)$-block-source. Therefore, following an argument similar to the one used above, the algorithm's declaration in this case is correct on some pair of corresponding subsources. ∎

In this section we gained some familiarity with the challenge-response mechanism and with the three-types lemma (Lemma 2.2), which is an important supplement to the mechanism. Hopefully, this experience will assist the reader in what follows.

3. Overview of the construction and analysis. In this section we present our construction of subextractors and give a comprehensive and detailed overview of the proof, though we allow ourselves to be somewhat imprecise whenever this contributes to the presentation. The formal proof, which can be recovered by the content of this section, appears in section 8. In section 3.1, we introduce the notions of entropy-trees and tree-structured sources. A variant of these notions was used by [6]. Then, in section 3.2, we overview the approach taken by [6] for their construction of two-source dispersers. Once the results needed from [6] are in place, in section 3.3 we give an overview for the rest of our construction and emphasize where our ideas deviate from theirs. In the following sections of this overview (see sections 3.4 and 3.5) we give further details.

3.1. Entropy-trees and tree-structured sources.

Motivating the notion of an entropy-tree. We already saw that an n-bit source with entropy-rate 0.6 has a subsource that is an $\Omega(n)$-block-source. Further, by applying the three-types lemma (Lemma 2.2) twice, we saw that either any source X with entropy-rate 0.4 has a subsource that is an $\Omega(n)$ block-source, or otherwise one of $\mathsf{left}(X), \mathsf{right}(X)$ is an $\Omega(n)$ block-source. We, however, are interested in sources X

with only polylog(n) entropy. Is it true that there is a block-source "lying somewhere" in X (or in a low-deficiency subsource of X) even for such low entropy? Yes, it is! Although we have to dig deeper.

To see why this is true, say X is an (n,k)-source. Lemma 2.2, set with $b=\sqrt{k}$, implies that there exists a deficiency $\sqrt{k}$ subsource X' of X. Either X' is a $\sqrt{k}$-block-source, or otherwise it holds that one of $\mathsf{left}(X'), \mathsf{right}(X')$ has almost all the entropy of X, namely, entropy $k-\sqrt{k}$. In other words, if X' is not a block-source, then the entropy-rate of one of $\mathsf{left}(X'), \mathsf{right}(X')$ has almost doubled.

Assume that X' is not a block-source, and that $\mathsf{left}(X')$ has entropy $k-\sqrt{k}$. By Lemma 2.2, set again with $b=\sqrt{k}$, there exists a deficiency $\sqrt{k}$ subsource $X''\subset X'$ such that either $\mathsf{left}(X'')$ is a $\sqrt{k}$-block-source, or otherwise one of $\mathsf{left}(\mathsf{left}(X''))$, $\mathsf{right}(\mathsf{left}(X''))$ has min-entropy $k-2\sqrt{k}$. That is, if $\mathsf{left}(X'')$ is also not a $\sqrt{k}$-block-source, then the entropy-rate of one of $\mathsf{left}(\mathsf{left}(X'')), \mathsf{right}(\mathsf{left}(X''))$ is almost four times the original entropy-rate of X.

Continuing this process, at some point we are bound to find a block-source. Indeed, if we failed to find a block-source in the first r iterations then there is a deficiency $r\sqrt{k}$ subsource $X^{(r)}\subset X$ and a length $n\cdot 2^{-r}$ block of $X^{(r)}$ that has entropy $k-r\sqrt{k}$. Thus, if $k-r\sqrt{k}>0.6n\cdot 2^{-r}$, then this block has a subsource that is a block-source, which will be found in the next iteration. Hence, if $k=\omega(\log^2 n)$, then a block-source will be found in the first $\log(n/k)+O(1)$ iterations of this process.[2]

As we apply Lemma 2.2 at most $\log n$ times in the process described above, and since in each application we move to a deficiency $\sqrt{k}$ subsource, we conclude that every (n,k)-source has a deficiency $\sqrt{k}\log n$ subsource that contains a block-source as a block. This block-source can be found (in the analysis) by following a certain "path of entropy" that determines which of the two halves of the current block of the source contains essentially all the entropy.

Entropy-trees. The above discussion naturally leads to what we call an *entropy-tree* and sources that have a tree-structure. An entropy-tree (see Figure 1) is a complete rooted binary tree T, where some of its nodes are labeled by one of the following labels: $\mathsf{H},\mathsf{B},\mathsf{F}$, which stand for high entropy, block-source, and fixed, respectively. The nodes of an entropy-tree are labeled according to rules that capture any possible entropy structure of a subsource obtained by the process described above. The rules are as follows:

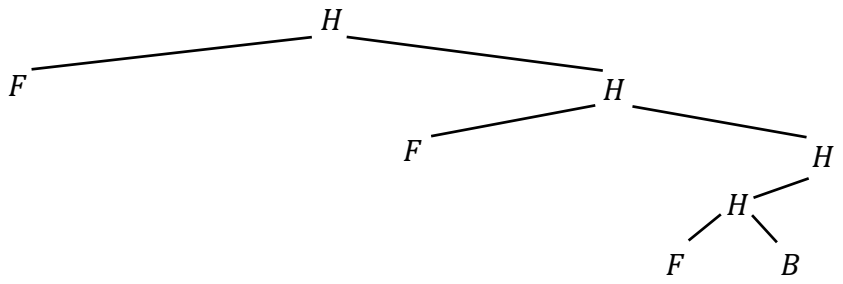

FIG. 1. *An example of an entropy-tree. Unlabeled nodes and edges to them do not appear in the figure.*

[2]Here and throughout the paper, the logarithm is always taken base 2.

- The root of T, denoted by $\mathsf{root}(T)$, is labeled by either H or B, expressing the fact that we assume the source itself has high entropy, compared to the entropy of the original source, and may even be a block-source.
- There is exactly one node in T, denoted by $v_{\mathsf{B}}(T)$, that is labeled by B. This expresses the fact we proved, namely, if one digs deep enough, a block-source will be found. The uniqueness of the node labeled by B captures the fact that we terminate the process once a block-source is found.
- If v is a nonleaf that has no label, or is otherwise labeled by F or B, then its sons have no label. This rule captures the fact that a node has no label if we are not interested in the block of the source that is associated with this node. Thus, if a block is fixed, we do not try to look for a block-source inside it. Similarly, if the node is a block-source, we stop the search.
- If v is a nonleaf that is labeled by H, then the sons of v can only be labeled according to the following rules:
  - If $\mathsf{leftSon}(v)$ is labeled by F, then $\mathsf{rightSon}(v)$ is labeled by either H or B.
  - If $\mathsf{leftSon}(v)$ is labeled by either H or B, then $\mathsf{rightSon}(v)$ has no label.

  These rules capture the guarantee of Lemma 2.2.

The entropy-path. With every entropy-tree T, we associate a path that we call the *entropy-path* of T. This is the unique path from $\mathsf{root}(T)$ to $v_{\mathsf{B}}(T)$. We say that a path in T contains the entropy-path if it starts at $\mathsf{root}(T)$ and goes through $v_{\mathsf{B}}(T)$. Note that we allow an entropy-tree to have nodes that are descendants of $v_{\mathsf{B}}(T)$. We just do not allow these nodes to have labels.

Tree-structured sources. Now that we have defined entropy-trees, we can say what it means for a source to have a T-structure, for some entropy-tree T. To this end we need to introduce some notation. Let n be an integer that is a power of 2. With a string $x \in \{0,1\}^n$, we associate a depth $\log n$ complete rooted binary tree, where with each node v of T we associate a substring x_v of x in the following natural way: $x_{\mathsf{root}(T)} = x$; and for $v \neq \mathsf{root}(T)$, if v is the left son of its parent, then $x_v = \mathsf{left}(x_{\mathsf{parent}(v)})$; otherwise, $x_v = \mathsf{right}(x_{\mathsf{parent}(v)})$.

Let T be a depth $\log n$ entropy-tree. An n-bit source X is said to have a T-structure with parameter k if for any node v in T the following hold:

- If v is labeled by F, then X_v is fixed to a constant.
- If v is labeled by H, then $\mathbf{H}_\infty(X_v) \geq k$.
- If v is labeled by B, then X_v is a $\sqrt{k}$-block-source.

With the notions of entropy-trees and tree-structured sources, we can summarize the discussion of this section by saying that any (n,k)-source, with $k = \omega(\log^2 n)$, has a deficiency $\sqrt{k}\log n$ subsource that has a T-structure with parameter $\Omega(k)$ for some entropy-tree T (that depends on the underlying distribution of X). Therefore, for the purpose of constructing subextractors, we may assume that we are given two independent samples from tree-structured sources rather than from general weak-sources.

One important observation to keep in mind is the following: By Fact 2.4 and by Lemma 2.6 it follows that if X' is a deficiency d subsource of a source having a T-structure with parameter $k = \omega(d)$, then X' has a T-structure with parameter $(1 - o(1))k$. In particular, we can move to $o(k)$ deficiency subsources throughout the analysis and still maintain the original tree-structure of the source.

3.2. Identifying the entropy-path. Tree-structured sources certainly seem nicer to work with than general weak-sources. However, it is still not clear what good this structure is if we do not have any information regarding the entropy-tree, and in

particular regarding the entropy-path.

Remarkably, by applying the challenge-response mechanism in a carefully chosen manner, Barak et al. [6] were able to identify the entropy-path of the entropy-tree T *algorithmically* given just one sample from $x \sim X$, where X is a T-structured source, and one sample from $y \sim Y$, where Y is a general weak-source that is independent of X. We now turn to describe the algorithm used by [6]. Before doing so, it is worth mentioning that Barak et al. proved something somewhat different. Indeed, they considered a variant of entropy-trees and had to prove a variant of what we need. In particular, their algorithm did not identify the entropy-path per se. Nevertheless, their proof can be adapted in a straightforward manner to obtain the result we describe next. We give a formal proof of what is needed for our construction in section 8.1.

What does it mean to identify the entropy-path? What do we mean by saying that an algorithm identifies the entropy-path of an entropy-tree T? This is an algorithm that on input $x, y \in \{0,1\}^n$, outputs a depth $\log n$ rooted complete binary tree and a marked root-to-leaf path on that tree, denoted by $p_{\mathsf{observed}}(x,y)$, the *observed* entropy-path. Ideally, the guarantee of the algorithm would have been the following: If x is sampled from a T-structured source X and y is sampled independently from a weak-source Y, then $p_{\mathsf{observed}}(x,y)$ contains the entropy-path of T with probability 1 over $(x,y) \sim (X,Y)$. That is, for any $(x,y) \in \mathsf{supp}(X,Y)$, if we draw the computed path $p_{\mathsf{observed}}(x,y)$ on the entropy-tree T, then this path starts at $\mathsf{root}(T)$ and goes through $v_{\mathsf{B}}(T)$.

Note that the path $p_{\mathsf{observed}}(x,y)$ is allowed to continue arbitrarily after visiting $v_{\mathsf{B}}(T)$. Asking that $p_{\mathsf{observed}}(x,y)$ will stop exactly at $v_{\mathsf{B}}(T)$ is a very strong requirement. In particular, it will conclude the construction of the subextractor. Indeed, once the block-source $X_{v_{\mathsf{B}}(T)}$ is found, one can simply output $\mathsf{BExt}(X_{v_{\mathsf{B}}(T)}, Y)$.

This was an ideal version of what we mean by identifying an entropy-path. For our needs, we will be satisfied with a weaker guarantee. Following [6], we will show that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$, such that with high probability over $(x,y) \sim (X',Y')$ it holds that $p_{\mathsf{observed}}(x,y)$ contains the entropy-path of T.

The fact that we only have a guarantee on low-deficiency subsources is good enough for us as we are aiming for a subextractor. The fact that there is an error (that did not appear in the analysis of [6]) should be handled with some care. Indeed, note that by moving to a deficiency d subsource, an ε error in the original source can grow to at most $2^d \cdot \varepsilon$ restricted to the subsource. We will make sure that the error is negligible compared to the deficiency we consider in the rest of the analysis. Thus, from here on we will suppress the error introduced in this step of identifying the entropy-path.

The algorithm of [6] for identifying the entropy-path. We now describe the algorithm that was devised by [6] for identifying the entropy-path of an entropy-tree T. The basic idea was depicted already in the toy algorithm from section 2.4. In fact, what the toy algorithm actually managed to do was identify the entropy-path of depth 2 tree-structured sources.

We first note that if $\mathsf{root}(T) = v_{\mathsf{B}}(T)$, then any observed entropy-path will contain $v_{\mathsf{B}}(T)$. So, we may assume that this is not the case. Let v be the parent of $v_{\mathsf{B}}(T)$ in T. As a first step, we want to determine which of the two sons of v is $v_{\mathsf{B}}(T)$. To this end, we will use the toy algorithm from section 2.4. More precisely, node v declares that its left son is $v_{\mathsf{B}}(T)$ if and only if

$$\mathsf{Response}\left(x_v, y, \mathsf{BExt}\left(x_{\mathsf{leftSon}(v)}, y\right)\right) = \mathsf{hasEntropy}. \tag{3.1}$$

Let us pause for a moment to introduce some notation. If (3.1) holds, we say that node v (x,y)-*favors its left-son*; otherwise, we say that v (x,y)-*favors its right son.* Moreover, we define the *good son* of v to be $v_{\mathsf{B}}(T)$. More generally, for a node $u \neq v_{\mathsf{B}}(T)$ that is an ancestor of $v_{\mathsf{B}}(T)$, we define the *good son* of u to be its unique son that is an ancestor of $v_{\mathsf{B}}(T)$. Note that by following the good sons from $\mathsf{root}(T)$ to $v_{\mathsf{B}}(T)$ one recovers the entropy-path of T. Thus, one correctly identifies the entropy-path of T on input x, y if and only if any ancestor of $v_{\mathsf{B}}(T)$ on the entropy-path of T (x,y)-favors its good son.

By following the proof of Claim 2.3, one can see that if $X_{\mathsf{leftSon}(v)}$ is fixed, then (3.1) holds with probability 0 on some low-deficiency subsources of X, Y. Further, by the challenge-response mechanism together with Fact 2.4 and Lemma 2.6, one can show that if $\mathsf{leftSon}(v) = v_{\mathsf{B}}(T)$, then with high probability over (X,Y), equation (3.1) holds. Observe that by the definition of an entropy-tree, these are the only two possible cases.

We showed how $v_{\mathsf{B}}(T)$ can "convince" its parent v that it is its good son. The trick was to use the block-source-ness of $X_{v_{\mathsf{B}}(T)}$ so as to generate a proper challenge. Considering one step further, we ask the following: If u is the parent of v, how can v convince u that it is its good son? After all, v is not a block-source. The elegant solution of Barak et al. is as follows. Given $x, y \in \{0,1\}^n$, the challenge of v will contain not only $\mathsf{BExt}(x_v, y)$ but also $\mathsf{BExt}(x_w, y)$, where w is v's (x,y)-favored son. Thus, if v's favored son happens to be its good son $v_{\mathsf{B}}(T)$, then the challenge posed by v will not be responded to by u.

More generally, a node v decides which of its two sons it (x,y)-favors not according to (3.1) but rather according to whether or not

$$\mathsf{Response}\left(x_v, y, \mathsf{GoodSonCh}\left(x_{\mathsf{leftSon}(v)}, y\right)\right) = \mathsf{hasEntropy}, \tag{3.2}$$

where $\mathsf{GoodSonCh}(x_{\mathsf{leftSon}(v)}, y)$ is a matrix with at most $\log n$ rows (according to the depth of the tree) that contains $\mathsf{BExt}(x_{\mathsf{leftSon}(v)}, y)$ as a row, as well as $\mathsf{BExt}(x_w, y)$, where w is the (x,y)-favored son of $\mathsf{leftSon}(v)$, and also $\mathsf{BExt}(x_r, y)$, where r is the (x,y)-favored son of w, etc.

The strategy of [6] for determining the output. Having found the entropy-path of T, we are in a much better shape. We know that one of the nodes on the path is a block-source. The trouble is that we still do not know which one. We conclude this section by saying only a few words about the strategy taken by [6] for resolving this problem, as at this point our strategy deviates from theirs. It is worth mentioning that the strategy taken by Barak et al. for determining the output is one place in the construction that poses a bottleneck for supporting entropy $2^{o(\sqrt{\log n})}$. It is also the reason why the number of output bits in their construction can be at most $O(\log \log n)$ and why the construction is only a disperser as apposed to a subextractor.

In order to output a nonconstant bit, as required by a 1 output bit disperser, Barak et al. assumed that the source X has some more structure. Not only should X have a T-structure, but it is also required that $\mathsf{left}(X_{v_{\mathsf{B}}(T)})$ have its own tree-structure. In particular, somewhere in the left block of $X_{v_{\mathsf{B}}(T)}$ there should be a second block-source. Note that this extra structure required from X can be assumed with almost no cost in parameters. Indeed, after applying the process from section 3.1 to X to obtain a deficiency $\sqrt{k}\log n$ subsource X' of X that has a tree-structure, one can simply apply the process again, this time to $\mathsf{left}(X'_{v_{\mathsf{B}}(T)})$, to find a second deficiency $\sqrt{k}\log n$ subsource of X with the desired structure.

Having this "double block-source" structure, Barak et al. were able to carefully tune the parameters of the challenge-response mechanism so that with some probability, $v_{\mathsf{B}}(T)$ will be convinced that $X_{\mathsf{leftSon}(v_{\mathsf{B}}(T))}$ contains a block-source, yet with some probability it will fail to notice that. With some more delicate work, and based on the fact that $X_{v_{\mathsf{B}}(T)}$ is a block-source, the fact that $v_{\mathsf{B}}(T)$'s decision is not constant can be carried upward all the way to $\mathsf{root}(T)$ and in turn can be translated to an output bit that is nonconstant.

3.3. The strategy for the rest of our construction. To carry the analysis of our subextractor, we require even more structure from our sources than the structure required by [6]. First, we require *both* X and Y to have a tree-structure. In previous works [5, 6], the second source Y was used mainly to "locate the entropy" of the source X, and the only assumption on Y was that it has a sufficient amount of entropy for this purpose. We, however, will make use of the structure of Y as well.

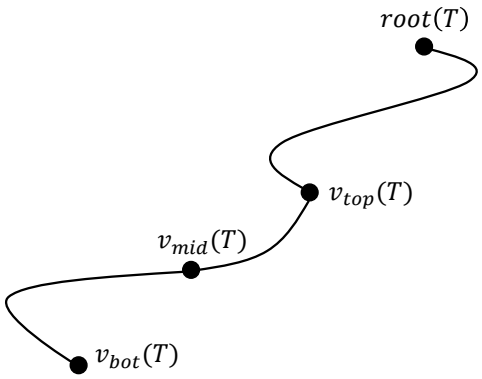

FIG. 2. *The "triple block-source" structure of an entropy-tree.*

Second, we need both X and Y to have a "triple block-source" structure (see Figure 2). That is, we assume that X has a T_X-structure with a node $v_{\mathsf{top}}(T_X)$ corresponding to the block-source $X_{v_{\mathsf{top}}(T_X)}$. We then assume that $\mathsf{left}(X_{v_{\mathsf{top}}(T_X)})$ has its own tree-structure with a node $v_{\mathsf{mid}}(T_X)$ corresponding to a second block-source $X_{v_{\mathsf{mid}}(T_X)}$ lying inside $\mathsf{left}(X_{v_{\mathsf{top}}(T_X)})$. Finally, we require that $\mathsf{left}(X_{v_{\mathsf{mid}}(T_X)})$ have its own tree-structure with a node $v_{\mathsf{bot}}(T_X)$ that corresponds to a third block-source $X_{v_{\mathsf{bot}}(T_X)}$ that lies inside $\mathsf{left}(X_{v_{\mathsf{mid}}(T_X)})$. The same goes for Y. Namely, Y also has a triple block-source structure. In particular, the entropy-tree of Y, denoted by T_Y, has nodes that we denote by $u_{\mathsf{top}}(T_Y)$, $u_{\mathsf{mid}}(T_Y)$, and $u_{\mathsf{bot}}(T_Y)$, analogous to $v_{\mathsf{top}}(T_X)$, $v_{\mathsf{mid}}(T_X)$, and $v_{\mathsf{bot}}(T_X)$ in T_X.

We allow ourselves to change the definition of an entropy-tree given in the previous section so that it will capture this "triple block-source" structure, but the reader should not worry about these details at this point. For the formal definition of entropy-trees and tree-structured sources, see section 6.

Given this structure of the sources, we are ready to give a high-level overview of our construction. In the subsequent sections of the overview (sections 3.4 and 3.5), we give further details. Let X be a T_X-structured source and let Y be a T_Y-structured source for some entropy-trees T_X, T_Y. At the first step, the subextractor identifies the entropy-path of T_X and the entropy-path of T_Y using the algorithm of [6]. More

precisely, given the samples $x \sim X$,$y \sim Y$, we compute two paths denoted by

$$p_{\mathsf{observed}}(x,y) = v_0(x,y), v_1(x,y), \ldots, v_{\log(n)-1}(x,y),$$
$$q_{\mathsf{observed}}(x,y) = u_0(x,y), u_1(x,y), \ldots, u_{\log(n)-1}(x,y).$$

This step must be done with some care. From technical reasons (related to the way the error term behaves when moving to subsources), we cannot use x, y to first find the entropy-path of T_X and then to find the entropy-path of T_Y. Thus, in some sense, the two paths must be computed simultaneously (see section 8.1 for more details).

At this point, ignoring some small error term, we have that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$, such that for any $(x,y) \in \mathsf{supp}(X',Y')$ it holds that $p_{\mathsf{observed}}(x,y)$ (resp., $q_{\mathsf{observed}}(x,y)$) contains the entropy-path of T_X (resp., T_Y). In particular, we have that $v_{\mathsf{depth}(v_{\mathsf{top}}(T_X))}(X',Y')$ is fixed to $v_{\mathsf{top}}(T_X)$, and the same holds for $v_{\mathsf{mid}}(T_X), v_{\mathsf{bot}}(T_X)$, as well as for $u_{\mathsf{top}}(T_Y)$, $u_{\mathsf{mid}}(T_Y)$, and $u_{\mathsf{bot}}(T_Y)$. To keep the notation clean, we write X, Y for X', Y' in this proof overview. That is, we assume that the entropy-paths are correctly identified on the tree-structured sources themselves.

At the second step of the algorithm, we identify $v_{\mathsf{mid}}(T_X)$ with high probability over subsources $X' \subset X$, $Y' \subset Y$. This sounds fantastic—having found $v_{\mathsf{mid}}(T_X)$, we can simply output $\mathsf{BExt}(X'_{v_{\mathsf{mid}}(T_X)}, Y')$ which is close to uniform. Unfortunately, however, the only way we know how to find $v_{\mathsf{mid}}(T_X)$ requires us to fix $\mathsf{left}(X'_{v_{\mathsf{mid}}(T_X)})$. That is, once found, $X'_{v_{\mathsf{mid}}(T_X)}$ is no longer a block-source.

We elaborate on how to find $v_{\mathsf{mid}}(T_X)$ in section 3.4. Then in section 3.5, we show how to determine the output of the subextractor even after losing the block-structure of $X_{v_{\mathsf{mid}}(T_X)}$.

3.4. Finding $v_{\mathsf{mid}}(T_X)$. Given $x, y \in \{0,1\}^n$, the key idea we use for identifying $v_{\mathsf{mid}}(T_X)$ on $p_{\mathsf{observed}}(x,y)$ lies in the design of a challenge that we call the *node-path challenge* (see Figure 3).

The node-path challenge and $v^{\mathsf{observed}}_{\mathsf{mid}}(x,y)$. Let v be a node in T_X, and let $q = w_0, \ldots, w_{\log(n)-1}$ be a root-to-leaf path in T_Y. We define the challenge $\mathsf{NodePathCh}(x_v, y_q)$ as the $\log(n)$-rows Boolean matrix such that for $i = 0, 1, \ldots, \log(n) - 1$,

$$\mathsf{NodePathCh}(x_v, y_q)_i = \mathsf{BExt}\left(y_{w_i}, x_v\right).$$

We define $v^{\mathsf{observed}}_{\mathsf{mid}}(x,y)$ to be the node v on $p_{\mathsf{observed}}(x,y)$ with the largest depth such that

$$\mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_v, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{hasEntropy}. \tag{3.3}$$

Informally speaking, based on the node-path challenge, a node on $p_{\mathsf{observed}}(x,y)$ uses the path $q_{\mathsf{observed}}(x,y)$ to prove that it is $v_{\mathsf{mid}}(T_X)$.

Ideally, we would want to prove that $v^{\mathsf{observed}}_{\mathsf{mid}}(x,y) = v_{\mathsf{mid}}(T_X)$ for any $(x,y) \in \mathsf{supp}(X,Y)$. By now we know that this is too much to ask, and in any case, it suffices to prove that there exist low-deficiency subsources $X' \subset X$, $Y' \subset Y$ such that with high probability over $(x,y) \sim (X',Y')$ it holds that $v^{\mathsf{observed}}_{\mathsf{mid}}(x,y) = v_{\mathsf{mid}}(T_X)$. Unfortunately, we will not be able to prove that either. What we will be able to show is that there exist strings α, β such that the following holds. Define

$$X_\alpha = X' \mid \left(X'_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))} = \alpha\right),$$
$$Y_\beta = Y' \mid \left(Y'_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))} = \beta\right),$$

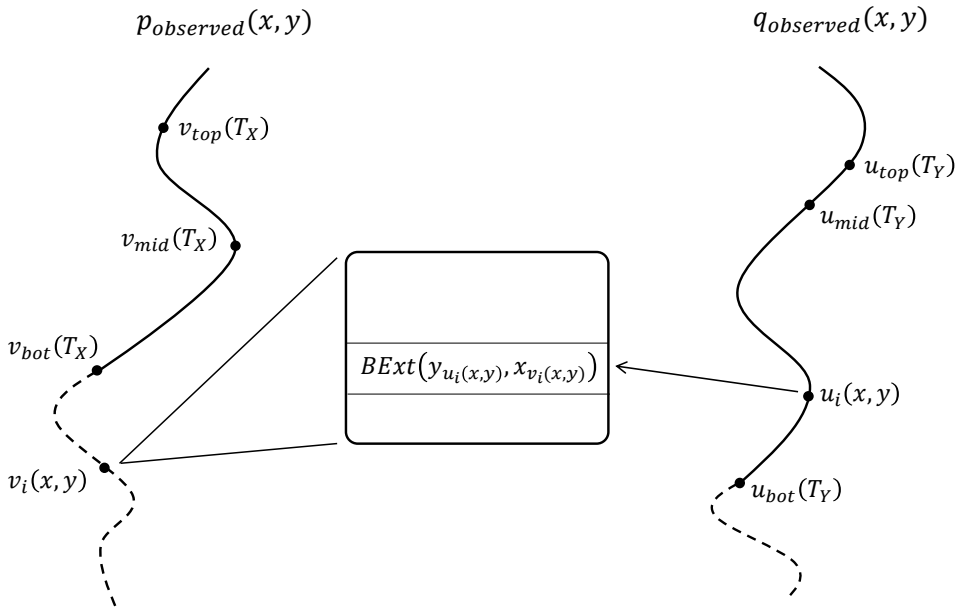

FIG. 3. *The node-path challenge.*

and let $i_{\mathsf{mid}}(T_X)$ denote the depth of $v_{\mathsf{mid}}(T_X)$.

The way we choose α, β is with respect to the error that we constantly ignore throughout this overview. Thus, assume that α, β are chosen in such a way that allows us to continue ignoring the error (this is done by a simple averaging argument). No further requirement is posed on α, β.

PROPOSITION 3.1. *There exist low-deficiency subsources $X_{\alpha,\beta} \subset X_\alpha$, $Y_{\alpha,\beta} \subset Y_\beta$, such that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, it holds that*

$$\forall i > i_{\mathsf{mid}}(T_X) \quad \mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_{v_i(x,y)}, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{fixed},$$
$$\mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_{v_{i_{\mathsf{mid}}(T_X)}(x,y)}, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{hasEntropy}.$$

Note that by the way we defined $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$, Proposition 3.1 yields that $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y) = v_{\mathsf{mid}}(T_X)$ with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$. In particular, this gives us an algorithm for computing $v_{\mathsf{mid}}(T_X)$—simply go up the computed path $p_{\mathsf{observed}}(x,y)$ until a node v is found for which (3.3) holds. In the rest of this section we prove Proposition 3.1.

The challenges of descendants of $v_{\mathsf{mid}}(T_X)$ on $p_{\mathsf{observed}}(x,y)$ are properly responded to. Proposition 3.1 has two parts. First, it states that the node-path challenges associated with nodes below $v_{i_{\mathsf{mid}}(T_X)}(x,y)$ on the path $p_{\mathsf{observed}}(x,y)$ are responded to with high probability over x, y that are sampled from some low-deficiency subsources of X_α, Y_β. Second, Proposition 3.1 states that the node-path challenge associated with $v_{i_{\mathsf{mid}}(T_X)}(x,y)$ is left unresponded to with high probability over the samples.

Recall that, ignoring a small error term, we assume that $v_{i_{\mathsf{mid}}(T_X)}(x,y) = v_{\mathsf{mid}}(T_X)$.

Let us first consider the nodes below $v_{\mathsf{mid}}(T_X)$ on $p_{\mathsf{observed}}(x,y)$. Naturally, we want to use the challenge-response mechanism. For that we must find low-deficiency subsources $X'_\alpha \subset X_\alpha$, $Y'_\beta \subset Y_\beta$ such that for all $i > i_{\mathsf{mid}}(T_X)$, the challenge

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha,Y'_\beta)}, (Y'_\beta)_{q_{\mathsf{observed}}(X'_\alpha,Y'_\beta)}\right) \tag{3.4}$$

is fixed to a constant. As was done in the analysis of the toy algorithm from section 2.4, to this end it is enough to show that the random variable

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha,Y_\beta)}, (Y_\beta)_{q_{\mathsf{observed}}(X'_\alpha,Y_\beta)}\right)$$

is a deterministic function of Y_β. Indeed, in such a case and since the challenge consists of a relatively small number of bits, we can apply Fact 2.5 to find a low-deficiency subsource $Y'_\beta \subset Y_\beta$ such that the random variable in (3.4) is fixed to a constant.

For $i > i_{\mathsf{mid}}(T_X)$, our starting point is the random variable

$$\mathsf{NodePathCh}\left((X_\alpha)_{v_i(X_\alpha,Y_\beta)}, (Y_\beta)_{q_{\mathsf{observed}}(X_\alpha,Y_\beta)}\right).$$

To make this random variable depend solely on Y_β, by moving to a subsource of X_α, we need to take care of all three appearances of X_α. We start with $q_{\mathsf{observed}}(X_\alpha, Y_\beta)$.

Claim 3.2. *There exists a deficiency* $\log n$ *subsource* $X'_\alpha \subset X_\alpha$ *such that* $q_{\mathsf{observed}}(X'_\alpha, Y_\beta)$ *is fixed to a constant.*

Proof. Let $i_{\mathsf{bot}}(T_Y)$ denote the depth of $u_{\mathsf{bot}}(T_Y)$. To prove the claim, we first recall that the path $q_{\mathsf{observed}}(X_\alpha, Y_\beta)$ contains the entropy-path of T_Y. In particular, we have that the nodes $u_0(X_\alpha,Y_\beta), \ldots, u_{i_{\mathsf{bot}}(T_Y)}(X_\alpha,Y_\beta)$ are fixed. It is left to argue that there is a low-deficiency subsource $X'_\alpha \subset X_\alpha$ such that the remaining nodes $u_{i_{\mathsf{bot}}(T_Y)+1}(X'_\alpha,Y_\beta), \ldots, u_{\log(n)-1}(X'_\alpha,Y_\beta)$ are fixed as well.

Let us first consider the random node $u_{i_{\mathsf{bot}}(T_Y)+1}(X_\alpha,Y_\beta)$ that is the son of the fixed node $u_{i_{\mathsf{bot}}(T_Y)}(X_\alpha,Y_\beta) = u_{\mathsf{bot}}(T_Y)$. According to (3.2), the node $u_{\mathsf{bot}}(T_Y)$ decides which of its two sons it favors, namely, which of its sons will be on $q_{\mathsf{observed}}(X_\alpha,Y_\beta)$, according to whether or not

$$\mathsf{Response}\left((Y_\beta)_{u_{\mathsf{bot}}(T_Y)}, X_\alpha, \mathsf{GoodSonCh}\left((Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))}, X_\alpha\right)\right) = \mathsf{hasEntropy}. \tag{3.5}$$

By the definition of an entropy-tree, $u_{\mathsf{bot}}(T_Y)$ is a descendant of $\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))$. Further, by definition, $(Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))}$ is fixed to β. Thus, also $(Y_\beta)_{u_{\mathsf{bot}}(T_Y)}$ and $(Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))}$ are fixed to some constants. Therefore, the Boolean expression in (3.5) is a deterministic function of X_α. By applying Fact 2.5, we obtain a deficiency 1 subsource X' of X_α such that the Boolean expression in (3.5) is fixed. In particular, $u_{i_{\mathsf{bot}}(T_Y)+1}(X',Y_\beta)$ is fixed to a constant.

At this point we can apply the same argument to $i_{\mathsf{bot}}(T_Y)+2$. Indeed, we have that $u_{i_{\mathsf{bot}}(T_Y)+1}(X',Y_\beta)$ is fixed to a constant and all appearances of Y_β in the Boolean expression that is analogous to (3.5) are again fixed to constants for the same reason as before. Since this process terminates after at most $\log n$ steps and since in each iteration we move to a deficiency 1 subsource of the previous obtained subsource, the claim follows. ∎

Given Claim 3.2, we turn to showing that for all $i > i_{\mathsf{mid}}(T_X)$,

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha,Y_\beta)}, (Y_\beta)_{q_{\mathsf{observed}}(X'_\alpha,Y_\beta)}\right) \tag{3.6}$$

is a deterministic function of Y_β. By the discussion above, this will prove the first part of Proposition 3.1.

By Claim 3.2, we already know that $q_{\mathsf{observed}}(X'_\alpha, Y_\beta)$ is fixed to a constant. Thus, it suffices to show that $(X'_\alpha)_{v_i(X'_\alpha,Y_\beta)}$ is a deterministic function of Y_β for all $i > i_{\mathsf{mid}}(T_X)$. By an argument similar to the one used in the proof of Claim 3.2, one can show that for any such i, $v_i(X'_\alpha, Y_\beta)$ is a deterministic function of Y_β. Note further that, by the definition of an entropy-tree, since $i > i_{\mathsf{mid}}(T_X)$, we have that $v_i(X'_\alpha, Y_\beta)$ is always (that is, for every $(x,y) \in \mathsf{supp}(X'_\alpha, Y_\beta)$) a descendant of $\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))$. Since $(X'_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ is fixed to a constant, we conclude that $(X'_\alpha)_{v_i(X'_\alpha,Y_\beta)}$ is indeed a deterministic function of Y_β.

By the discussion above, we are now in a position to apply Fact 2.5 to obtain a low-deficiency subsource $Y'_\beta \subset Y_\beta$ such that

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha,Y'_\beta)}, (Y'_\beta)_{q_{\mathsf{observed}}(X'_\alpha,Y'_\beta)}\right)$$

is fixed to a constant. We can then apply the challenge-response mechanism and conclude that there exist low-deficiency subsources $X_{\alpha,\beta} \subset X'_\alpha$, $Y_{\alpha,\beta} \subset Y'_\beta$ such that for any $(x,y) \in \mathsf{supp}((X_{\alpha,\beta}, Y_{\alpha,\beta}))$, it holds that

$$\forall i > i_{\mathsf{mid}}(T_X) \quad \mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_{v_i(x,y)}, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{fixed}.$$

The challenge of $v_{\mathsf{mid}}(T_X)$ is left unresponded to. To prove Proposition 3.1, it is left to show that the node-path challenge associated with $v_{\mathsf{mid}}(T_X)$ is unresponded to. More precisely, it suffices to show that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, it holds that

$$\mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_{v_{\mathsf{mid}}(T_X)}, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{hasEntropy}.$$

Since $u_{\mathsf{top}}(T_Y)$ is on the path $q_{\mathsf{observed}}(x,y)$ for all $(x,y) \in \mathsf{supp}(X_{\alpha,\beta}, Y_{\alpha,\beta})$, the matrix

$$\mathsf{NodePathCh}\left((X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}, (Y_{\alpha,\beta})_{q_{\mathsf{observed}}(X_{\alpha,\beta},Y_{\alpha,\beta})}\right)$$

contains the row

$$\mathsf{BExt}\left((Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}, (X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}\right). \tag{3.7}$$

Since $X_{v_{\mathsf{mid}}(T_X)}$ is a block-source, $(X_\alpha)_{v_{\mathsf{mid}}(T_X)}$ has a significant amount of entropy. Indeed, X_α is obtained from X by fixing $X_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))} = \mathsf{left}(X_{v_{\mathsf{mid}}(T_X)})$. Since $X_{\alpha,\beta}$ is a low-deficiency subsource of X_α, Fact 2.5 then implies that $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}$ also has a significant amount of entropy.

We now observe that $(Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}$ is a block-source. Indeed, $Y_{u_{\mathsf{top}}(T_Y)}$ is a block-source and Y_β is obtained from Y by fixing $Y_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))}$. Since $Y_{u_{\mathsf{mid}}(T_Y)}$ is a block-source, this fixing leaves some entropy in $(Y_\beta)_{u_{\mathsf{mid}}(T_Y)}$. Recall further that $(Y_\beta)_{u_{\mathsf{mid}}(T_Y)}$ lies inside $\mathsf{left}((Y_\beta)_{u_{\mathsf{top}}(T_Y)})$ as $u_{\mathsf{mid}}(T_Y)$ is a descendant of $\mathsf{leftSon}(u_{\mathsf{top}}(T_Y))$. Thus, $(Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}$ is a block-source.

Consider now any low-deficiency subsources $\hat{X} \subset X_{\alpha,\beta}$, $\hat{Y} \subset Y_{\alpha,\beta}$. By Fact 2.5 and by Lemma 2.6 we have that $\hat{X}_{v_{\mathsf{mid}}(T_X)}$ has a significant amount of entropy and that $\hat{Y}_{u_{\mathsf{top}}(T_Y)}$ is a block-source (with some deterioration in parameters). Thus, for any low-deficiency subsources $\hat{X}, \hat{Y}$ of $X_{\alpha,\beta}$, $Y_{\alpha,\beta}$, respectively, we have that the challenge matrix associated with $v_{\mathsf{mid}}(T_X)$ contains a row that is close to uniform. In particular, this matrix is close to having high entropy. Thus, by the challenge-response mechanism, we have that the node-path challenge associated with $v_{\mathsf{mid}}(T_X)$ is left unresponded to with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, as desired.

3.5. Determining the output. At the last step of the algorithm, we compute the output of the subextractor that is defined as

$$\mathsf{SubExt}(x,y) = \mathsf{BExt}\left(x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)} \circ x, y\right),$$

where by $x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)} \circ x$ we denote the block-source with the first block $x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)}$ and the second block that equals x. Technically, we need to append the first block with zeros so that both blocks will have the same length, and also append y with zeros, but we ignore such minor technicalities in this section.

There are two potential problems with applying BExt the way we do above. First, we see that the block-source fed to BExt depends on the sample y, which is problematic since y is used as a sample from the weak-source as well. This, however, is a nonissue. Indeed, recall that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$ it holds that $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y) = v_{\mathsf{mid}}(T_X)$, and so ignoring a small error, the computation of the extractor BExt above is the same as

$$\mathsf{BExt}\left(x_{v_{\mathsf{mid}}(T_X)} \circ x, y\right).$$

Now that we have shown that there are no dependencies between the two samples fed to BExt, we only need to make sure that the first sample is indeed coming from a block-source when sampling $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$.

To see why this is true, recall that $v_{\mathsf{mid}}(T_X)$ is a descendant of $\mathsf{leftSon}(v_{\mathsf{top}}(T_X))$ and that $X_{v_{\mathsf{top}}(T_X)}$ is a block-source. Since $X_{\alpha,\beta}$ is obtained from X by fixing $X_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ (and by moving to low-deficiency subsources) and since $X_{v_{\mathsf{mid}}(T_X)}$ is a block-source, we have that $(X_{\alpha,\beta})_{v_{\mathsf{top}}(T_X)}$ is also a block-source. Therefore, $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)} \circ X_{\alpha,\beta}$ is also a block-source. This shows that the application of BExt above is valid, and that the output is close to uniform with high probability over $(X_{\alpha,\beta}, Y_{\alpha,\beta})$.

4. Preliminaries.

4.1. Standard (and less standard) notations and definitions. The logarithm in this paper is always taken base 2. For every natural number $n \geq 1$, define $[n] = \{1, 2, \ldots, n\}$.

Strings and matrices. Let n be an even integer. Let $x \in \{0,1\}^n$. For $i \in [n]$, we let x_i denote the ith bit of x. For $\emptyset \neq I \subseteq [n]$, we let x_I denote the projection of x to the coordinate set I. That is, if $I = \{i_1, \ldots, i_m\}$ with $i_1 < i_2 < \cdots < i_m$ then $x_I = x_{i_1} x_{i_2} \cdots x_{i_m}$. We denote by $\mathsf{left}(x)$ the $n/2$ leftmost bits of x and by $\mathsf{right}(x)$ the $n/2$ rightmost bits of x. That is, $\mathsf{left}(x) = x_1 \cdots x_{n/2}$ and $\mathsf{right}(x) = x_{(n/2)+1} \cdots x_n$. We denote the concatenation of two strings x, y by $x \circ y$. The length of x is denoted by $|x|$. Given an $r \times n$ matrix x, for $i = 0, 1, \ldots, r-1$, we let x_i denote row i of x. Note that we start the row numbering from 0.

Trees. Let T be a complete rooted binary tree. We denote the root of T by $\mathsf{root}(T)$. Throughout the paper we consider trees where some of the nodes are labeled by labels from a ground set L. If v is a labeled node in a tree T, we denote its label by $\mathsf{label}(v)$. If v is a nonleaf in T, we denote the left and right sons of v by $\mathsf{leftSon}(v), \mathsf{rightSon}(v)$, respectively. If v is not the root of T, $\mathsf{parent}(v)$ denotes the (unique) parent of v. The depth of T is denoted by $\mathsf{depth}(T)$. The depth of a node v in T, denoted by $\mathsf{depth}(v)$, is the distance in edges from $\mathsf{root}(T)$ to v. Note that $\mathsf{depth}(\mathsf{root}(T)) = 0$.

Random variables and distributions. We sometimes abuse notation and syntactically treat a random variable and its distribution as equal. Let X, Y be two random variables. We say that Y is a *deterministic function of* X if the value of X determines the value of Y. Namely, there exists a (deterministic) function f such that $Y = f(X)$. Throughout the paper, we mostly use capital letters to denote random variables.

Associating strings with trees. Let n be a power of 2 and let $x \in \{0,1\}^n$. The tree that is associated with x, denoted by T_x, is a depth $\log n$ complete rooted binary tree, where with each node v of T_x we associate a substring x_v of x as follows:

- $x_{\mathsf{root}(T)} = x$;
- For $v \neq \mathsf{root}(T)$, if v is the left son of its parent, then $x_v = \mathsf{left}\left(x_{\mathsf{parent}(v)}\right)$; otherwise, $x_v = \mathsf{right}\left(x_{\mathsf{parent}(v)}\right)$.

Statistical distance. The *statistical distance* between two distributions X, Y on a common domain D is defined by

$$\mathsf{SD}\left(X, Y\right) = \max_{A \subseteq D} \left\{ \left| \mathbf{Pr}[X \in A] - \mathbf{Pr}[Y \in A] \right| \right\}.$$

If $\mathsf{SD}(X, Y) \leq \varepsilon$ we say that X is ε-close to Y and write $X \sim_\varepsilon Y$.

Min-entropy. The *min-entropy* of a random variable X is defined by

$$\mathbf{H}_\infty(X) = \min_{x \in \mathsf{supp}(X)} \log\left(\frac{1}{\mathbf{Pr}[X = x]}\right).$$

If X is supported on $\{0,1\}^n$, we define the *min-entropy rate* of X by $\mathbf{H}_\infty(X)/n$. In such a case, if X has min-entropy k or more, we say that X is an (n,k)-weak-source or simply an (n,k)-source.

In some cases we will consider a random variable X that is ε-close, in statistical distance, to some k-source Y, though X itself might have very low entropy. In such case we sometimes say that X is ε-close to having min-entropy k and write $\mathbf{H}^\varepsilon_\infty(X) \geq k$. This notion is sometimes referred to as *smooth min-entropy* in the literature (see, e.g., [39]).

4.2. Li's block-source–weak-source extractor. Let X be a random variable on n bit strings, and assume n is even. We say that X is an (n,k)-*block-source* if the following hold:

- $\mathbf{H}_\infty(\mathsf{left}(X)) \geq k$.
- For any $x \in \mathsf{supp}(\mathsf{left}(X))$ it holds that $\mathbf{H}_\infty(\mathsf{right}(X) \mid \mathsf{left}(X) = x) \geq k$.

We sometimes omit the length n of X and say that X is a k-block-source.

In a recent breakthrough, Li [30] gave a construction of an extractor for two n-bit sources, where the first source is a polylog(n)-block-source and the second is a weak-source with min-entropy polylog(n). Our construction heavily relies on Li's extractor.

THEOREM 4.1 (see [30]). *There exists a universal constant $\gamma > 0$ such that the following holds. For all integers n, k with $k \geq \log^{12} n$, there is a* poly(n)*-time computable function*

$$\mathsf{BExt}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$$

such that if X is a k-block-source, where each block is on $n/2$ bits, and Y is an independent (n,k)-source, then

$$\mathsf{SD}\left((\mathsf{BExt}(X,Y), Y), (U_m, Y)\right) \leq \varepsilon,$$

and

$$\mathsf{SD}\left((\mathsf{BExt}(X,Y),X),(U_m,X)\right) \leq \varepsilon,$$

where $m = 0.9k$ *and* $\varepsilon = 2^{-k^\gamma}$.

We remark that using results for [14], one can improve the required entropy $k = \log^{12} n$ in Theorem 4.1 to $\widetilde{\Omega}(\log^7 n)$.

Let $t < n$ be even integers. We sometimes apply BExt on strings $x \in \{0,1\}^t$ and $y \in \{0,1\}^n$ and write $\mathsf{BExt}(x,y)$. Formally, we actually compute $\mathsf{BExt}(x',y)$ where x' is obtained by adding $(n-t)/2$ zeros before and after x. This way of padding x preserves the block-structure of x.

4.3. Subsources. The notion of a subsource was first explicitly introduced and studied by Barak et al. [5]. We start by giving the definition of a subsource and then collect some facts about subsources.

DEFINITION 4.2 (subsource). *Given random variables* X *and* X' *on* $\{0,1\}^n$*, we say that* X' *is a* deficiency d *subsource of* X *and write* $X' \subset X$ *if there exists a set* $A \subseteq \{0,1\}^n$ *such that* $X' = X \mid (X \in A)$ *and* $\mathbf{Pr}[X \in A] \geq 2^{-d}$*. More precisely, for every* $a \in A$*,* $\mathbf{Pr}[X' = a]$ *is defined by* $\mathbf{Pr}[X = a \mid X \in A]$*, and for* $a \notin A$*,* $\mathbf{Pr}[X' = a] = 0$.

We make frequent use of the following simple facts about subsources.

FACT 4.3 ([6], Fact 3.11). *If* X *is an* (n,k)*-source and* X' *is a deficiency* d *subsource of* X*, then* X' *is an* $(n, k-d)$*-source.*

FACT 4.4 ([6], Fact 3.13). *Let* X *be a random variable on* n*-bit strings. Let* $f\colon \{0,1\}^n \to \{0,1\}^\ell$ *be a function. Then, there exist* $c \in \{0,1\}^\ell$ *and a deficiency* ℓ *subsource* $X' \subset X$ *such that* $f(x) = c$ *for every* $x \in \mathsf{supp}(X')$.

LEMMA 4.5 (see [6, Lemma 3.15]). *Let* X *be a random variable that is* ε*-close to having min-entropy* k*, with* $\varepsilon < 1/4$*. Then, there exists a deficiency* 2 *subsource* $X' \subset X$ *that has min-entropy* $k-3$.

LEMMA 4.6 (see [6, Corollary 3.19]). *Let* X *be a* k*-block-source, and let* X' *be a deficiency* d *subsource of* X*. Then,* X' *is* ε*-close to a* $k - d - \log(1/\varepsilon) - 1$ *block-source.*

5. The challenge-response mechanism. In this section we further abstract the challenge-response mechanism that was introduced in [5] and refined by [6]. This abstraction will make it easier for us to apply the mechanism in our proofs. The reader is referred to section 2 for an intuitive-level overview of the challenge-response mechanism.

THEOREM 5.1. *For integers* $\ell < n$*, there exists a* $\mathrm{poly}(n)$*-time computable function*

$$\mathsf{Resp}\colon \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^\ell \to \{\mathsf{fixed}, \mathsf{hasEntropy}\}$$

with the following property. For any two independent n*-bit sources* X, Y *that are* $1/4$*-close to having min-entropy* $\Omega(\log^{10} n)$*, and for any function* $\mathsf{Challenge}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^\ell$*, the following hold:*

- *If* $\mathsf{Challenge}(X,Y)$ *is fixed to a constant, then there exist deficiency* $2\ell + 2$ *subsources* $X' \subset X$*,* $Y' \subset Y$*, such that*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X',Y')}\left[\mathsf{Resp}\left(x,y,\mathsf{Challenge}(x,y)\right) = \mathsf{fixed}\right] = 1.$$

- *If for any deficiency* 20ℓ *subsources* $\hat{X} \subset X$, $\hat{Y} \subset Y$ *it holds that* $\mathsf{Challenge}(\hat{X}, \hat{Y})$ *is* ε*-close to having min-entropy* k*, then*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X,Y)}[\mathsf{Resp}\,(x, y, \mathsf{Challenge}(x, y)) = \mathsf{fixed}] \leq (2^{-k} + \varepsilon) \cdot \mathrm{poly}(n).$$

For the proof of Theorem 5.1 we make use of the following theorem.

THEOREM 5.2 (see [6, Theorem 4.3]). *There exist universal constants* γ, c *such that for any integer* n*, there exists a* $\mathrm{poly}(n)$*-time computable function*

$$\mathsf{SE}\colon \{0,1\}^n \times \{0,1\}^n \to \left(\{0,1\}^\ell\right)^r,$$

with $\ell \leq \gamma k$ *and* $r = n^c$*, such that the following holds. For any two independent* $(n, \log^{10} n)$*-sources* X, Y*, the following hold.*

- *Let* c *be any fixed* ℓ *bit string. Then, there exist subsources* $X_c \subset_{2\ell} X$, $Y_c \subset_{2\ell} Y$ *and an index* $i \in [r]$ *such that* $\mathbf{Pr}\,[\mathsf{SE}(X_c, Y_c)_i = c] = 1$.
- *Given any particular row index* $i \in [r]$, (X, Y) *is* $2^{-10\ell}$*-close to a convex combination of subsources such that for every* $(\hat{X}, \hat{Y})$ *in the combination it holds that*
  - $\hat{X}$ *is a deficiency* 20ℓ *subsource of* X,
  - $\hat{Y}$ *is a deficiency* 20ℓ *subsource of* Y,
  - $\hat{X}, \hat{Y}$ *are independent, and*
  - $\mathsf{SE}(\hat{X}, \hat{Y})_i$ *is fixed to a constant.*

Proof of Theorem 5.1. We start by describing the algorithm for computing the response function $\mathsf{Response}(x, y, \mathsf{Challenge}(x, y))$ as described in [5]. The algorithm computes $\mathsf{SE}(x, y)$, where the output length of SE is set to ℓ. The algorithm then checks whether or not $\mathsf{Challenge}(x, y)$ appears as a row in $\mathsf{SE}(x, y)$. If so, the algorithm outputs fixed; otherwise it outputs $\mathsf{hasEntropy}$.

We turn to the analysis. Assume first that $\mathsf{Challenge}(X, Y)$ is fixed to some constant c. By Lemma 4.5, there exist deficiency 2 subsources $X' \subset X$, $Y' \subset Y$ such that X', Y' have min-entropy $\Omega(\log^{10} n)$. Note that $\mathsf{Challenge}(X', Y')$ is also fixed to c. By Theorem 5.2, there exist deficiency 2ℓ subsources $X_c \subset X'$, $Y_c \subset Y'$ and an index $i \in [r]$ such that with probability 1 over $(x, y) \sim (X_c, Y_c)$ it holds that $\mathsf{Challenge}(x, y) = \mathsf{SE}(x, y)_i$, thus proving the first part of the theorem.

Assume now that for any deficiency 20ℓ subsources $\hat{X} \subset X$, $\hat{Y} \subset Y$, it holds that $\mathsf{Challenge}(\hat{X}, \hat{Y})$ is ε-close to having min-entropy k. Consider any fixed $i \in [r]$. By Theorem 5.2, (X, Y) is $2^{-10\ell}$-close to a convex combination of subsources such that every $(\hat{X}, \hat{Y})$ in the combination has the four listed properties. Since $\mathsf{Challenge}(\hat{X}, \hat{Y})$ is ε-close to having min-entropy k and since $\mathsf{SE}(\hat{X}, \hat{Y})_i$ is fixed, we have that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(\hat{X},\hat{Y})}[\mathsf{Challenge}(x, y) = \mathsf{SE}(x, y)_i] \leq 2^{-k} + \varepsilon.$$

Accounting for the distance from (X, Y) to the convex combination,

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X,Y)}[\mathsf{Challenge}(x, y) = \mathsf{SE}(x, y)_i] \leq 2^{-k} + \varepsilon + 2^{-10\ell}.$$

Therefore, by the union bound over all $i \in [r]$,

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X,Y)}[\exists i \in [r] \quad \mathsf{Challenge}(x, y) = \mathsf{SE}(x, y)_i] \leq (2^{-k} + \varepsilon + 2^{-10\ell})r.$$

As $r = \mathrm{poly}(n)$ and since $k \leq \ell$, the proof follows. ∎

6. Entropy-trees and tree-structured sources. In this section we define entropy-trees and tree-structured sources. These are variants of notions that were introduced in [6]. Then, in Proposition 6.5, we show that any weak-source has a low-deficiency subsource that is a tree-structured source.

DEFINITION 6.1 (entropy-trees). *An* entropy-tree T *is a complete rooted binary tree where some of the nodes of the tree are labeled by one of the following labels:* $\mathsf{F}, \mathsf{H}, \mathsf{B}_{\mathsf{top}}, \mathsf{B}_{\mathsf{mid}}, \mathsf{B}_{\mathsf{bot}}$*, according to the following set of rules:*

- $\mathsf{label}(\mathsf{root}(T)) \in \{\mathsf{H}, \mathsf{B}_{\mathsf{top}}\}$.
- *There is exactly one node in* T *that is labeled by* $\mathsf{B}_{\mathsf{top}}$*, one node that is labeled by* $\mathsf{B}_{\mathsf{mid}}$*, and one node labeled by* $\mathsf{B}_{\mathsf{bot}}$*, denoted by* $v_{\mathsf{top}}(T)$*,* $v_{\mathsf{mid}}(T)$*, and* $v_{\mathsf{bot}}(T)$*, respectively. Further,* $v_{\mathsf{mid}}(T)$ *is a (possibly immediate) descendant of* $\mathsf{leftSon}(v_{\mathsf{top}}(T))$*, and* $v_{\mathsf{bot}}(T)$ *is a (possibly immediate) descendant of* $\mathsf{leftSon}(v_{\mathsf{mid}}(T))$*. We denote* $i_{\mathsf{top}}(T) = \mathsf{depth}(v_{\mathsf{top}}(T))$*,* $i_{\mathsf{mid}}(T) = \mathsf{depth}(v_{\mathsf{mid}}(T))$*, and* $i_{\mathsf{bot}}(T) = \mathsf{depth}(v_{\mathsf{bot}}(T))$*.*
- *If* v *is a nonleaf that has no label or otherwise is labeled by* F *or* $\mathsf{B}_{\mathsf{bot}}$*, then both its sons have no label.*
- *If* v *is a nonleaf labeled by* H*, then* $\mathsf{leftSon}(v)$ *has a label. Further,*
  - *If* $\mathsf{label}(\mathsf{leftSon}(v)) = \mathsf{F}$*, then* $\mathsf{rightSon}(v)$ *has a label different than* F*.*
  - *If* $\mathsf{label}(\mathsf{leftSon}(v)) \neq \mathsf{F}$*, then the right son of* v *has no label.*
- *If* v *is a nonleaf labeled by* $\mathsf{B}_{\mathsf{top}}$ *or* $\mathsf{B}_{\mathsf{mid}}$*, then* $\mathsf{leftSon}(v)$ *has a label. Further,*
  - *If* $\mathsf{label}(\mathsf{leftSon}(v)) = \mathsf{F}$*, then* $\mathsf{label}(\mathsf{rightSon}(v)) = \mathsf{H}$*.*
  - *If* $\mathsf{label}(\mathsf{leftSon}(v)) \neq \mathsf{F}$*, then the right son of* v *has no label.*

DEFINITION 6.2 (entropy-path). *Let* T *be an entropy-tree. The* entropy-path *of* T *is the path that starts at* $\mathsf{root}(T)$ *and ends at* $v_{\mathsf{bot}}(T)$*. We denote the nodes on this path by* $\mathsf{root}(T) = v_0(T), v_1(T), \ldots, v_{i_{\mathsf{bot}}(T)}(T) = v_{\mathsf{bot}}(T)$*. We say that a path* p *in* T contains *the entropy-path of* T *if* p *starts at* $\mathsf{root}(T)$ *and goes through* $v_{\mathsf{bot}}(T)$*.*

In our proofs we oftentimes consider two sources X, Y, each having its own tree-structure. For ease of reading, we use v to denote a node in one entropy-tree and u to denote a node in the other entropy-tree. For example, say X has a T_X-structure and Y has a T_Y-structure, for some entropy trees T_X, T_Y. Then, the entropy-path of T_X is denoted by $v_0(T_X), v_1(T_X), \ldots, v_{i_{\mathsf{bot}}(T_X)}(T_X) = v_{\mathsf{bot}}(T_X)$, whereas $u_0(T_Y), u_1(T_Y), \ldots, u_{i_{\mathsf{bot}}(T_Y)}(T_Y) = u_{\mathsf{bot}}(T_Y)$ is the entropy-path of the entropy-tree T_Y.

DEFINITION 6.3 (good son). *Let* T *be an entropy-tree and let* $v \neq v_{\mathsf{bot}}(T)$ *be an ancestor of* $v_{\mathsf{bot}}(T)$*. The* good son *of* v *is defined to be the unique son of* v *that is an ancestor of* $v_{\mathsf{bot}}(T)$*.*

We note that the entropy-path of an entropy-tree T is the path obtained by following the good son of each node starting from $\mathsf{root}(T)$ until reaching $v_{\mathsf{bot}}(T)$.

DEFINITION 6.4 (tree-structured sources). *Let* T *be an entropy-tree. We say that an* n*-bit random variable* X *has a* T-structure *with parameters* (k, ε) *if the following hold. For any node* v *in* T*,*

- *If* $\mathsf{label}(v) = \mathsf{F}$*, then* X_v *is fixed to a constant.*
- *If* $\mathsf{label}(v) = \mathsf{H}$*, then the following hold:*
  - *If* v *is an ancestor of* $v_{\mathsf{top}}(T)$*, then* $\mathbf{H}^{\varepsilon}_{\infty}(X_v) \geq k$*.*
  - *If* v *is a descendant of* $v_{\mathsf{top}}(T)$ *and an ancestor of* $v_{\mathsf{mid}}(T)$*, then* $\mathbf{H}^{\varepsilon}_{\infty}(X_v) \geq \sqrt{k}$*.*
  - *If* v *is a descendant of* $v_{\mathsf{mid}}(T)$*, then* $\mathbf{H}^{\varepsilon}_{\infty}(X_v) \geq k^{1/4}$*.*

- $X_{v_{\mathsf{top}}}$ *is* ε*-close to a* $\sqrt{k}$*-block-source.*
- $X_{v_{\mathsf{mid}}}$ *is* ε*-close to a* $k^{1/4}$*-block-source.*
- $X_{v_{\mathsf{bot}}}$ *is* ε*-close to a* $k^{1/8}$*-block-source.*

By a simple counting argument one can show that most weak-sources do not have a tree-structure (at least not with nontrivial parameters). Nevertheless, in the following proposition we show that any weak-source has a low-deficiency subsource that has a tree-structure. A similar statement appears in Lemma 6.10 of [6].

PROPOSITION 6.5. *Let* X *be an* (n,k)*-source with* $k=\omega(\log^8 n)$*. Then, there exists a deficiency* $O(\sqrt{k}\log n)$ *subsource of* X *that has a* T*-structure, for some entropy-tree* T*, with parameters* $(\Omega(k), 2^{-\Omega(k^{1/4})})$.

In the rest of this section we prove Proposition 6.5. We start by giving a proof for a special case of the "fixing entropies lemma" by Barak et al. [6] (see their Lemma 3.20), which will be sufficient for our needs.

LEMMA 6.6. *Let* X *be an* (n,k)*-source. Let* $0<\tau_1<\tau_2<n$ *be any two numbers. Set* $\tau_0=0$ *and* $\tau_3=n$*. Then, there exist a deficiency* 2 *subsource* $X'\subset X$ *and an index* $i\in\{0,1,2\}$ *such that the following hold:*

- *For any* $c\in\mathsf{supp}(\mathsf{left}(X'))$, $\mathbf{H}_\infty(\mathsf{right}(X')\mid\mathsf{left}(X')=c)\in[\tau_i,\tau_{i+1}]$.
- $\mathbf{H}_\infty(\mathsf{left}(X'))+\tau_{i+1}\geq k-2$.

Proof. Define a function $f\colon\mathsf{supp}(\mathsf{left}(X))\to\{0,1,2\}$ as follows:

$$f(c)=\begin{cases}0, & \mathbf{H}_\infty(\mathsf{right}(X)\mid\mathsf{left}(X)=c)\in[\tau_0,\tau_1);\\ 1, & \mathbf{H}_\infty(\mathsf{right}(X)\mid\mathsf{left}(X)=c)\in[\tau_1,\tau_2);\\ 2, & \mathbf{H}_\infty(\mathsf{right}(X)\mid\mathsf{left}(X)=c)\in[\tau_2,\tau_3].\end{cases}$$

By Fact 4.4 (where we identify the range $\{0,1,2\}$ with an arbitrary subset of $\{0,1\}^2$), there exists a deficiency 2 subsource $X'\subset X$ for which $f(X')$ is fixed to some constant $i\in\{0,1,2\}$. The first property of the lemma readily follows.

As for the second item, let $t=\mathbf{H}_\infty(\mathsf{left}(X'))$. Then, there exists $x\in\mathsf{supp}(X')$ such that $\mathbf{Pr}[X'=x]\geq 2^{-(t+\tau_{i+1})}$, and so $\mathbf{H}_\infty(X')\leq t+\tau_{i+1}$. On the other hand, since X' is a deficiency 2 subsource of X, Fact 4.3 implies that $\mathbf{H}_\infty(X')\geq\mathbf{H}_\infty(X)-2\geq k-2$, which concludes the proof. ∎

Next we prove a lemma that is analogous to the two-types lemma by Barak et al. (see [6, Lemma 6.8]).

LEMMA 6.7 (three-types lemma). *For any* (n,k)*-source* X *and an integer* $b<k/2-1$*, there exists a deficiency* $b+2$ *subsource* $X'\subset X$ *such that (at least) one of the following holds:*

- X' *is a* b*-block-source.*
- $\mathbf{H}_\infty(\mathsf{left}(X'))\geq k-b-2$.
- $\mathsf{left}(X')$ *is fixed to a constant and* $\mathbf{H}_\infty(\mathsf{right}(X'))\geq k-b-2$.

Proof. Set $\tau_1=b$, $\tau_2=k-b-2$, and note that by our assumption on b, $\tau_1<\tau_2$. Apply Lemma 6.6 to obtain a deficiency 2 subsource $X''\subset X$ and $i\in\{0,1,2\}$. We consider three cases, according to the value of i.

- If $i=0$, then by the second item of Lemma 6.6, $\mathbf{H}_\infty(\mathsf{left}(X''))+\tau_1\geq k-2$. Thus, $\mathbf{H}_\infty(\mathsf{left}(X''))\geq k-b-2$. We then take $X'=X''$ to conclude the proof.
- If $i=1$, then for any fixing of $\mathsf{left}(X'')$, we have that $\mathbf{H}_\infty(\mathsf{right}(X''))\in[\tau_1,\tau_2]=[b,k-b-2]$. By the second item of Lemma 6.6, $\mathbf{H}_\infty(\mathsf{left}(X''))\geq$

$k-2-\tau_2 = b$. Therefore, X'' is a b-block-source, and we take $X' = X''$ to conclude the proof.

- By Lemma 6.6, if $i = 2$, then for any fixing of $\mathsf{left}(X'')$, we have that $\mathbf{H}_\infty(\mathsf{right}(X'')) \geq \tau_2 = k - b - 2$. Let x be an element that maximizes the probability of the event $\{\mathsf{left}(X'') = x\}$. Let $X' = X'' \mid \{\mathsf{left}(X'') = x\}$ be a subsource of X''. This costs another b in deficiency. ∎

We are now ready to prove Proposition 6.5.

Proof of Proposition 6.5. Let T be a depth $\log n$ complete rooted binary tree. During the proof, we label some of the nodes of T, following the rules of entropy-trees, such that eventually we will find a low-deficiency subsource of X that has a T-structure. We start by applying Lemma 6.7 with $b = \sqrt{k}$ to obtain a deficiency $b+2$ subsource $X' \subset X$ for which one of the following holds:

- X' is a b-block-source.
- $\mathbf{H}_\infty(\mathsf{left}(X')) \geq k - b - 2$.
- $\mathsf{left}(X')$ is fixed to a constant and $\mathbf{H}_\infty(\mathsf{right}(X')) \geq k - b - 2$.

If the first case occurs, we label $\mathsf{root}(T)$ by $\mathsf{B_{top}}$. Otherwise, if the second case occurs, we label $\mathsf{root}(T)$ by H and continue recursively to the source $\mathsf{left}(X')$ with the tree rooted at $\mathsf{leftSon}(\mathsf{root}(T))$. Otherwise, we are guaranteed that the third case happens. We then label $\mathsf{root}(T)$ by H and $\mathsf{leftSon}(\mathsf{root}(T))$ by F and continue recursively to the source $\mathsf{right}(X')$ with the tree rooted at $\mathsf{rightSon}(\mathsf{root}(T))$. We continue in this manner until a node is labeled by $\mathsf{B_{top}}$.

Note that after $r-1$ recursive calls occur without labeling any node by $\mathsf{B_{top}}$, the rth recursive call considers an (n', k')-source, with $n' = n \cdot 2^{-r}$ and $k' \geq k - r(b+2)$. Note further that if $k' - b - 2 > 0.5n'$, the only case that can occur in Lemma 6.7 is the first case, in which case the process above will label a node by $\mathsf{B_{top}}$. In particular, by our choice $b = \sqrt{k}$, the process can continue for at most $\log(n/k) + 2$ recursive calls. Thus, a deficiency $\sqrt{k} \cdot \log n$ subsource $X' \subset X$ will be found and a node of T will be labeled by $\mathsf{B_{top}}$, such that the block of X' that corresponds to this node is a $\sqrt{k}$-block-source. We denote this node by v_{top}; namely, $X'_{v_{\mathsf{top}}}$ is a $\sqrt{k}$-block-source.

We now apply the same process to $\mathsf{left}(X'_{v_{\mathsf{top}}})$ with $b' = k^{1/4}$. This time, once a b'-block-source is found, we label the corresponding node v_{mid} by $\mathsf{B_{mid}}$. Using an argument similar to the one used above, and since $\mathsf{left}(X'_{v_{\mathsf{top}}})$ has min-entropy at least $b = \sqrt{k}$, one can show that as $k = \omega(\log^4 n)$ there exists a deficiency $k^{1/4} \log n$ subsource $X'' \subset X'$ such that $X''_{v_{\mathsf{mid}}}$ is a b'-block-source.

Finally, we apply the process to $\mathsf{left}(X''_{v_{\mathsf{mid}}})$ with $b'' = k^{1/8}$. Once a b''-block-source is found, we label the corresponding node v_{bot} by $\mathsf{B_{bot}}$. One can show that there exists a deficiency $k^{1/8} \log n$ subsource $X''' \subset X''$ such that $X'''_{v_{\mathsf{bot}}}$ is a b''-block-source.

For every node v that is labeled by H, which is an ancestor of v_{top}, it holds that $\mathbf{H}_\infty(X_v) \geq k - (b+2) \cdot 2 \log n \geq k/2$. Since X''' is a deficiency $O(\sqrt{k} \log n)$ subsource of X, Fact 4.3 implies that $\mathbf{H}_\infty(X'''_v) \geq k/3$. Similarly, for every node v that is labeled by H, which is a descendant of v_{top} and an ancestor of v_{mid}, it holds that $\mathbf{H}_\infty(X'''_v) \geq b - (b'+2)\log n - 6 \log n = \Omega(\sqrt{k})$. Further, for any node v labeled by H that is a descendant of v_{mid} and an ancestor of v_{bot}, $\mathbf{H}_\infty(X'''_v) \geq b' - b'' \log n = \Omega(k^{1/4})$.

As for the block-sources, recall that $X_{v_{\mathsf{top}}}$ is a $b = \sqrt{k}$-block-source. Since X''' is a deficiency $O(k^{1/4} \log n)$ subsource of X, Lemma 4.6 implies that $X'''_{v_{\mathsf{top}}}$ is $2^{-\Omega(\sqrt{k})}$-close to an $\Omega(\sqrt{k})$-block-source. Similarly, $X'''_{v_{\mathsf{mid}}}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source. Finally, $X'''_{v_{\mathsf{bot}}}$ is an $\Omega(k^{1/8})$-block-source. ∎

7. The two-source subextractor. In this section we describe our two-source subextractor. Let n be a power of 2, and let ℓ be a parameter. On input $x, y \in \{0,1\}^n$, the computation of the subextractor is done in three steps.

Step 1: Identifying the entropy-paths. Informally speaking, the goal of the first step of the algorithm is to identify the entropy-paths of the entropy-trees from which the samples x, y were presumably sampled. This step is a variant of a component from the two-source disperser by [6]. For this step we make use of the challenge-response mechanism. Thus, algorithmically, we start by setting up suitable challenges.

Setting the good son challenges. Recall that with the n-bit strings x, y we associate depth $\log n$ complete rooted binary trees that are denoted by T_x, T_y, respectively (see section 4). With each node v of T_x, we associate a $\log(n) \times \ell$ Boolean matrix, denoted by $\mathsf{GoodSonCh}(x_v, y)$, computed from leaves to root, recursively, as follows. All the entries in rows $0, \ldots, \mathsf{depth}(v)-1$ of $\mathsf{GoodSonCh}(x_v, y)$ are fixed to 0. Row $\mathsf{depth}(v)$ of $\mathsf{GoodSonCh}(x_v, y)$ contains $\mathsf{BExt}(x_v, y)$, where BExt is the extractor from Theorem 4.1 set to output ℓ bits. If v is a nonleaf, rows $\mathsf{depth}(v)+1, \ldots, \log(n)-1$ are copied from the respective rows of $\mathsf{GoodSonCh}(x_{\mathsf{leftSon}(v)}, y)$ or from the respective rows of $\mathsf{GoodSonCh}(x_{\mathsf{rightSon}(v)}, y)$ according to the following rule. If

$$\mathsf{Resp}\left(x_v, y, \mathsf{GoodSonCh}\left(x_{\mathsf{leftSon}(v)}, y\right)\right) = \mathsf{fixed}$$

then rows $\mathsf{depth}(v)+1, \ldots, \log(n)-1$ of $\mathsf{GoodSonCh}(x_v, y)$ are taken from the corresponding rows of $\mathsf{GoodSonCh}(x_{\mathsf{rightSon}(v)}, y)$. Otherwise, these rows are taken from the corresponding rows of $\mathsf{GoodSonCh}(x_{\mathsf{leftSon}(v)}, y)$. In the first case we say that v (x,y)-*favors its right son*, and in the second case we say that v (x,y)-*favors its left son*.

Analogously, with each node u of T_y we associate a $\log(n) \times \ell$ Boolean matrix, denoted by $\mathsf{GoodSonCh}(y_u, x)$, defined recursively as follows. All entries in rows $0, \ldots, \mathsf{depth}(u)-1$ of $\mathsf{GoodSonCh}(y_u, x)$ are fixed to 0. Row $\mathsf{depth}(u)$ of $\mathsf{GoodSonCh}(y_u, x)$ contains $\mathsf{BExt}(y_u, x)$. If u is a nonleaf, rows $\mathsf{depth}(u)+1, \ldots, \log(n)-1$ are copied from the respective rows of $\mathsf{GoodSonCh}(y_{\mathsf{leftSon}(u)}, x)$ or from the respective rows of $\mathsf{GoodSonCh}(y_{\mathsf{rightSon}(u)}, x)$ according to the following rule. If

$$\mathsf{Resp}\left(y_u, x, \mathsf{GoodSonCh}\left(y_{\mathsf{leftSon}(u)}, x\right)\right) = \mathsf{fixed}$$

then the remaining rows are taken from the corresponding rows of $\mathsf{GoodSonCh}(y_{\mathsf{rightSon}(u)}, x)$. Otherwise, the rows are taken from the corresponding rows of $\mathsf{GoodSonCh}(y_{\mathsf{leftSon}(u)}, x)$. In the first case we say that u (x,y)-*favors its right son*, and in the second case we say that u (x,y)-*favors its left son*.

Computing the entropy-paths. Let $v_0(x,y), v_1(x,y), \ldots, v_{\log(n)-1}(x,y)$ be the root-to-leaf path in T_x such that $v_i(x,y)$ (x,y)-favors $v_{i+1}(x,y)$ for all $i = 0, 1, \ldots, \log(n)-2$. Similarly, let $u_0(x,y), u_1(x,y), \ldots, u_{\log(n)-1}(x,y)$ be the root-to-leaf path in T_y such that $u_i(x,y)$ (x,y)-favors $u_{i+1}(x,y)$ for all $i = 0, 1, \ldots, \log(n)-2$. We denote $v_0(x,y), \ldots, v_{\log(n)-1}(x,y)$ by $p_{\mathsf{observed}}(x,y)$ and call this path the *observed entropy-path of* T_x. Similarly, we denote the path $u_0(x,y), \ldots, u_{\log(n)-1}(x,y)$ by $q_{\mathsf{observed}}(x,y)$ and call this path the *observed entropy-path of* T_y.

The computation done in Step 1. Given $x, y \in \{0,1\}^n$, in Step 1 the subextractor computes $p_{\mathsf{observed}}(x,y)$ and $q_{\mathsf{observed}}(x,y)$. Note that this computation can be done in poly(n)-time.

Step 2: Identify $v_{\mathsf{mid}}(T_X)$. Given $x, y, p_{\mathsf{observed}}(x,y)$ and $q_{\mathsf{observed}}(x,y)$, at the second step the algorithm computes a function we denote by $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$. To this end we make a second use of the challenge-response mechanism. Thus, algorithmically, we start by setting suitable challenges.

The node-path challenges. Let $\ell' < \ell$ be a parameter. Let v be a node in T_x and let $p = w_0, \ldots, w_{\log(n)-1}$ be a root-to-leaf path in T_y. The node-path challenge associated with (v,p), that we denote by $\mathsf{NodePathCh}(x_v, y_p)$, is a $\log(n) \times \ell'$ Boolean matrix defined as follows. For $j = 0, \ldots, \log(n) - 1$,

$$\mathsf{NodePathCh}(x_v, y_p)_j = \mathsf{BExt}(y_{w_j}, x_v),$$

where BExt is the extractor from Theorem 4.1 set to output ℓ' bits.

Computing $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$. We define $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$ to be the node v in $p_{\mathsf{observed}}(x,y)$ with the largest depth such that

$$\mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_v, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{hasEntropy}.$$

If no such node exists we define v, arbitrarily, as $\mathsf{root}(T_X)$. Note that $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$ can be computed in poly(n)-time.

Step 3: Determining the output. Given x, y, and $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$ which was computed in the previous step, the output of the subextractor is defined by

$$\mathsf{SubExt}(x,y) = \mathsf{BExt}\left(x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)} \circ x, y\right),$$

where $x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)} \circ x$ is the block-source with first block $x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)}$ and second block equal to x. Technically, we need to append $x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)}$ with zeros to obtain a length $|x| = n$ string. Similarly, we append y with n zeros to obtain a $2n$-bit string.

Recap. We conclude this section by summarizing the three high-level steps in the computation of the subextractor. On input $x, y \in \{0,1\}^n$:

1. Compute $p_{\mathsf{observed}}(x,y)$ and $q_{\mathsf{observed}}(x,y)$.
2. Compute $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$.
3. Output $\mathsf{BExt}(x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)} \circ x, y)$.

8. Analysis of the construction. In this section we prove Theorem 1.10 by analyzing the algorithm from section 7. Recall that the algorithm is parameterized by two parameters, denoted by ℓ and ℓ'. For an error parameter ε we set $\ell' = c_1 \cdot \log(n/\varepsilon)$ for some suitable constant c_1. We further set $\ell = c_2 \ell' \cdot \log^2(n)$ for some suitable constant c_2. With this choice of parameters we prove the following theorem which readily implies Theorem 1.10.

THEOREM 8.1 (two-source subextractors). *Let γ be the constant from Theorem* 4.1*. Then, there exists a constant c such that the algorithm described in section* 7 *with the choice of ℓ, ℓ' above is a subextractor for outer-entropy $k_{\mathsf{out}} = (\log(n/\varepsilon))^{O(1/\gamma)}$ and inner-entropy $k_{\mathsf{in}} = \Omega(k_{\mathsf{out}}^{1/4})$ with $m = \Omega(k_{\mathsf{out}}^{1/4})$ output bits and error ε.*

For ease of notation, throughout this section we denote k_{out} by k. The proof of Theorem 8.1 is done in three steps, following the three steps of the construction from section 7. By Proposition 6.5, we may assume that X has a T_X-structure and that Y has a T_Y-structure for some entropy-trees T_X, T_Y, each with parameters $(\Omega(k), 2^{-\Omega(k^{1/4})})$. This costs only $O(\sqrt{k} \log n)$ in deficiency.

8.1. Analysis of Step 1. We start this section by proving the following claim.

CLAIM 8.2. *There exist deficiency $\ell \log^2 n$ subsources $X_{\mathsf{F}} \subset X$, $Y_{\mathsf{F}} \subset Y$ with the following property. For every node v in T_X that is labeled by F it holds that $\mathsf{GoodSonCh}((X_{\mathsf{F}})_v, Y_{\mathsf{F}})$ is fixed to a constant. Further, for every node u in T_Y that is labeled by F it holds that $\mathsf{GoodSonCh}((Y_{\mathsf{F}})_u, X_{\mathsf{F}})$ is fixed to a constant.*

Proof. Let v be a node in T_X that is labeled by F. Since X has a T_X-structure, X_v is fixed to a constant, and so $\mathsf{GoodSonCh}(X_v, Y)$ is a deterministic function of Y. Since $\mathsf{GoodSonCh}(X_v, Y)$ consists of $\ell \log n$ bits, Fact 4.4 implies that there exists a deficiency $\ell \log n$ subsource $Y' \subset Y$ such that $\mathsf{GoodSonCh}(X_v, Y')$ is fixed to a constant. Repeating this argument for every $v \in T_X$ that is labeled by F, we get a subsource $Y_{\mathsf{F}} \subset Y$ such that $\mathsf{GoodSonCh}(X_v, Y_{\mathsf{F}})$ is fixed to a constant for every v in T_X that is labeled by F. By the definition of an entropy-tree, there is at most one node labeled by F in each level of T_X and since $\mathsf{depth}(T_X) = \log n$, we have that Y_{F} is a deficiency $\ell \log^2 n$ subsource of Y.

As Y_{F} is a subsource of Y, for every node u in T_Y that is labeled by F it holds that $(Y_{\mathsf{F}})_u$ is fixed to a constant. We now perform the analogous process on T_Y to obtain a deficiency $\ell \log^2 n$ subsource $X_{\mathsf{F}} \subset X$ such that for every node u in T_Y that is labeled by F it holds that $\mathsf{GoodSonCh}((Y_{\mathsf{F}})_u, X_{\mathsf{F}})$ is fixed to a constant. Note that since X_{F} is a subsource of X, it also holds that $\mathsf{GoodSonCh}((X_{\mathsf{F}})_v, Y_{\mathsf{F}})$ is fixed to a constant for every v in T_X that is labeled by F. Thus, informally speaking, by performing the analogous process to T_Y we have not "ruined" the desired property we obtained first for T_X. ∎

Next we show that there exist low-deficiency subsources $X_{\mathsf{FI}} \subset X_{\mathsf{F}}$, $Y_{\mathsf{FI}} \subset Y_{\mathsf{F}}$ (FI stands for "fixed identified"), restricted to which, the algorithm correctly identifies the nodes in T_X, T_Y that are labeled by F.

CLAIM 8.3. *There exist deficiency $O(\ell \log^2 n)$ subsources $X_{\mathsf{FI}} \subset X_{\mathsf{F}}$, $Y_{\mathsf{FI}} \subset Y_{\mathsf{F}}$ with the following property. For every node v of T_X that is labeled by F and for every node u of T_Y that is labeled by F, it holds that*

$$\mathbf{Pr}\left[\mathsf{parent}(v) \ (X_{\mathsf{FI}}, Y_{\mathsf{FI}})\text{-}favors\ v\right] = 0,$$
$$\mathbf{Pr}\left[\mathsf{parent}(u) \ (X_{\mathsf{FI}}, Y_{\mathsf{FI}})\text{-}favors\ u\right] = 0.$$

Proof. Let v be a node in T_X that is labeled by F. We first note that by the definition of an entropy-tree, $\mathsf{root}(T_X)$ cannot be labeled by F, and so it is valid to refer to $\mathsf{parent}(v)$. Further, by the definition of an entropy-tree, if a node is labeled by F then it must be the left son of its parent. Hence, $\mathsf{parent}(v)$ (x, y)-favors v if and only if

$$\mathsf{Response}\left(x_{\mathsf{parent}(v)}, y, \mathsf{GoodSonCh}\left(x_{\mathsf{leftSon}(\mathsf{parent}(v))}, y\right)\right) = \mathsf{hasEntropy}. \tag{8.1}$$

By Claim 8.2, $\mathsf{GoodSonCh}((X_{\mathsf{F}})_v, Y_{\mathsf{F}}) = \mathsf{GoodSonCh}((X_{\mathsf{F}})_{\mathsf{leftSon}(\mathsf{parent}(v))}, Y_{\mathsf{F}})$ is fixed to a constant. Thus, to apply the challenge-response mechanism to conclude that (8.1) holds with probability 0 restricted to low-deficiency subsources of $X_{\mathsf{F}}, Y_{\mathsf{F}}$, we only need to show that both $(X_{\mathsf{F}})_{\mathsf{parent}(v)}$ and Y_{F} have a sufficient amount of entropy. We start with Y_{F}. As Y_{F} is a deficiency $\ell \log^2 n$ subsource of Y and since $\mathbf{H}_\infty(Y) = \Omega(k)$, our choice of ℓ together with Fact 4.3 implies that $\mathbf{H}_\infty(Y_{\mathsf{F}}) = \Omega(k)$.

As for the entropy of $(X_{\mathsf{F}})_{\mathsf{parent}(v)}$, by the definition of entropy-trees and tree-structured sources, since $\mathsf{label}(v) = \mathsf{F}$ it holds that $\mathsf{label}(\mathsf{parent}(v)) \in \{\mathsf{H}, \mathsf{B}_{\mathsf{top}}, \mathsf{B}_{\mathsf{mid}}\}$ and so $X_{\mathsf{parent}(v)}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$. Since X_{F} is a

deficiency $\ell \log^2 n$ subsource of X we have that $(X_{\mathsf{F}})_{\mathsf{parent}(v)}$ is $2^{-\Omega(k^{1/4})+\ell \log^2 n}$-close to having min-entropy $\Omega(k^{1/4}) - \ell \log^2 n$. By our choice of ℓ, we have that the expression $\Omega(k^{1/4}) - \ell \log^2 n$ in the entropy and in the error term is $\Omega(k^{1/4})$, and so $(X_{\mathsf{F}})_{\mathsf{parent}(v)}$ is $2^{-\Omega(k/\log^2 n)}$-close to having min-entropy $\Omega(k^{1/4})$.

As $2^{-\Omega(k^{1/4})} \leq 1/4$, Theorem 5.1 implies that there exist deficiency $2\ell \log(n) + 2$ subsources $X' \subset X_{\mathsf{F}}$, $Y' \subset Y_{\mathsf{F}}$ such that for any $(x, y) \in \mathsf{supp}((X', Y'))$, equation (8.1) fails to hold. Thus, for any such (x, y) it holds that $\mathsf{parent}(v)$ does not (x, y)-favor v.

We repeat this argument for every node v in T_X that is labeled by F and obtain deficiency $(2\ell \log(n) + 2) \cdot \log n = O(\ell \log^2 n)$ subsources $X'' \subset X_{\mathsf{F}}$, $Y'' \subset Y_{\mathsf{F}}$ with the property that for every v in T_X that is labeled by F it holds that

$$\mathbf{Pr}\left[\mathsf{parent}(v) \ (X'', Y'')\text{-favors } v\right] = 0.$$

We can repeat the process above in such a way since the entropy of Y remains large enough throughout the process, and furthermore, for all v labeled by one of $\{\mathsf{H}, \mathsf{B}_{\mathsf{top}}, \mathsf{B}_{\mathsf{mid}}\}$ it holds that X_v remains close to having high min-entropy.

We now apply the same argument for every node u in T_Y that is labeled by F. Since X'' and Y'' are deficiency $O(\ell \log^2 n)$ subsources of $X_{\mathsf{F}}, Y_{\mathsf{F}}$, respectively, we can obtain deficiency $O(\ell \log^2 n)$ subsources $X_{\mathsf{FI}} \subset X''$, $Y_{\mathsf{FI}} \subset Y''$, such that for any node u in T_Y that is labeled by F it holds that

$$\mathbf{Pr}\left[\mathsf{parent}(u) \ (X_{\mathsf{FI}}, Y_{\mathsf{FI}})\text{-favors } u\right] = 0.$$

We note that since X_{FI} and Y_{FI} are subsources of X'', Y'', it also holds that for every node v in T_X that is labeled by F,

$$\mathbf{Pr}\left[\mathsf{parent}(v) \ (X_{\mathsf{FI}}, Y_{\mathsf{FI}})\text{-favors } v\right] = 0.$$

That is, we have not "ruined" the desired property we obtained first in T_X when working on T_Y. This concludes the proof of the claim. ∎

Up to this point, we found deficiency $O(\ell \log^2 n)$ subsources $X_{\mathsf{FI}} \subset X$ and $Y_{\mathsf{FI}} \subset Y$ such that the nodes labeled by F in T_X, T_Y are correctly identified by the challenge-response mechanism when applied to samples from $X_{\mathsf{FI}}, Y_{\mathsf{FI}}$. Next we prove that with high probability over $(X_{\mathsf{FI}}, Y_{\mathsf{FI}})$, the entropy-paths in T_X, T_Y are identified correctly by the subextractor in the sense that the observed entropy-paths contain the entropy-paths of the respective entropy-trees.

CLAIM 8.4. *Except with probability* $2^{-\Omega(\ell)}$ *over* $(x, y) \sim (X_{\mathsf{FI}}, Y_{\mathsf{FI}})$*, it holds that*

$$\forall i \in \{0, \ldots, i_{\mathsf{bot}}(T_X)\} \quad v_i(x, y) = v_i(T_X),$$
$$\forall i \in \{0, \ldots, i_{\mathsf{bot}}(T_Y)\} \quad u_i(x, y) = u_i(T_Y).$$

Proof. We prove the first equation in the statement of the claim. The proof of the second equation is similar, and then the proof of the claim follows by the union bound. We first observe that by the definition of an entropy-tree, for any ancestor $v \neq v_{\mathsf{bot}}(T_X)$ of $v_{\mathsf{bot}}(T_X)$ it holds that $\mathsf{label}(\mathsf{leftSon}(v)) = \mathsf{F}$ if and only if $\mathsf{rightSon}(v)$ is the good son of v. Indeed, on one hand, if $\mathsf{leftSon}(v)$ is labeled by F then $\mathsf{leftSon}(v)$ cannot be an ancestor of $v_{\mathsf{bot}}(T_X)$ as all of $\mathsf{leftSon}(v)$'s descendants have no label. On the other hand, since v has a label and its label can only be one of $\mathsf{H}, \mathsf{B}_{\mathsf{top}}, \mathsf{B}_{\mathsf{mid}}$, if its left son is not labeled by F then $\mathsf{rightSon}(v)$ has no label, and so $\mathsf{rightSon}(v)$ cannot be an ancestor of v as all of its descendants have no label.

Ideally, given this observation, we would have liked to prove by a backward induction on $i = i_{\mathsf{bot}}(T_X) - 1, \ldots, 1, 0$ that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X_{\mathsf{Fl}},Y_{\mathsf{Fl}})}[\forall j \in \{i, \ldots, i_{\mathsf{bot}}(T_X) - 1\} \quad v_j(T_X) \ (x,y)\text{-favors its good son}] \geq 1 - 2^{-\Omega(\ell)}.$$

Indeed, note that the claim will then follow by considering $i = 0$. However, we need to prove a stronger statement to have a stronger induction hypothesis, as otherwise we will not be able to carry out the induction step. More precisely, set $t = 20\ell \log n$. Let $\varepsilon_{i_{\mathsf{bot}}(T_X)-1} = 2^{-\Omega(\ell)}$. For $i = i_{\mathsf{bot}}(T_X) - 2, \ldots, 1, 0$, define $\varepsilon_i = (2^{-\Omega(\ell)} + \varepsilon_{i+1}) \cdot \mathrm{poly}(n)$. We prove by a backward induction on $i = i_{\mathsf{bot}}(T_X) - 1, \ldots, 1, 0$ that for any deficiency $i \cdot t$ subsources $X' \subset X_{\mathsf{Fl}}$, $Y' \subset Y_{\mathsf{Fl}}$, it holds that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X',Y')}[\forall j \in \{i, \ldots, i_{\mathsf{bot}}(T_X) - 1\} \quad v_j(T_X) \ (x,y)\text{-favors its good son}] \geq 1 - \varepsilon_i.$$

We note that the claim follows by considering $i = 0$ as $\varepsilon_0 = 2^{-\Omega(\ell)} \cdot 2^{O(\log^2 n)} = 2^{-\Omega(\ell)}$.

We start with the base of the induction $i = i_{\mathsf{bot}}(T_X) - 1$. Let $X' \subset X_{\mathsf{Fl}}$, $Y' \subset Y_{\mathsf{Fl}}$ be deficiency $(i_{\mathsf{bot}}(T_X) - 1) \cdot t$ subsources. Consider two cases according to the label of $\mathsf{leftSon}(v_{i_{\mathsf{bot}}(T_X)-1}(T_X))$. If $\mathsf{label}(\mathsf{leftSon}(v_{i_{\mathsf{bot}}(T_X)-1}(T_X))) = \mathsf{F}$ then by Claim 8.3,

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X_{\mathsf{Fl}},Y_{\mathsf{Fl}})}\left[v_{i_{\mathsf{bot}}(T_X)-1}(T_X) \ (x,y)\text{-favors its left son}\right] = 0.$$

Since X', Y' are subsources of $X_{\mathsf{Fl}}, Y_{\mathsf{Fl}}$, respectively, the same holds for $(x,y) \sim (X', Y')$. Moreover, as the good son of $v_{i_{\mathsf{bot}}(T_X)-1}(T_X)$ is its right son, the basis of the induction for this case follows.

Consider now the case $\mathsf{label}(\mathsf{leftSon}(v_{i_{\mathsf{bot}}(T_X)-1}(T_X))) \neq \mathsf{F}$. By the observation above, in this case, the good son of $v_{i_{\mathsf{bot}}(T_X)-1}(T_X)$ is its left son, and so $\mathsf{leftSon}(v_{i_{\mathsf{bot}}(T_X)-1}(T_X)) = v_{\mathsf{bot}}(T_X)$. Thus, $v_{i_{\mathsf{bot}}(T_X)-1}(T_X)$ (x,y)-favors its good son if and only if

$$\mathsf{Response}\left(x_{v_{i_{\mathsf{bot}}(T_X)-1}(T_X)}, y, \mathsf{GoodSonCh}\left(x_{v_{\mathsf{bot}}(T_X)}, y\right)\right) = \mathsf{hasEntropy}. \tag{8.2}$$

To conclude the proof of the base case, it is enough to show that (8.2) holds with probability $1 - 2^{-\Omega(\ell)}$ over $(x,y) \sim (X', Y')$. To this end, recall that $\mathsf{GoodSonCh}(x_{v_{\mathsf{bot}}(T_X)}, y)$ contains $\mathsf{BExt}(x_{v_{\mathsf{bot}}(T_X)}, y)$ as a row. By Theorem 5.1, it is enough to show that for all deficiency t subsources $\hat{X} \subset X'$, $\hat{Y} \subset Y'$, it holds that $\mathsf{BExt}(\hat{X}_{v_{\mathsf{bot}}(T_X)}, \hat{Y})$ is close to uniform.

By Lemma 4.6, applied with $\delta = 2^{-\Omega(k^{1/8})}$, since $\hat{X}$ is a deficiency $i_{\mathsf{bot}}(T_X) \cdot t + O(\ell \log^2 n) = O(\ell \log^2 n)$ subsource of X and since $X_{v_{\mathsf{bot}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/8})$-block-source, $\hat{X}_{v_{\mathsf{bot}}(T_X)}$ is $2^{-\Omega(k^{1/8})}$-close to an $\Omega(k^{1/8})$-block-source. Note that we used the fact that

$$\ell \cdot \log^2 n = c_2 \ell' \cdot \log^4 n = c_1 c_2 \log(n/\varepsilon) \cdot \log^4 n = O\left(k^{1/8}\right).$$

Now, as $\hat{Y}$ is a deficiency $O(\ell \log^2 n)$ subsource of Y, and since $\mathbf{H}_\infty(Y) \geq k$, $\mathbf{H}_\infty(\hat{Y}) = \Omega(k)$. Since $k^{1/8} = \Omega(\log^8 n)$, Theorem 4.1 and the remark following it imply that $\mathsf{BExt}(\hat{X}_{v_{\mathsf{bot}}(T_X)}, \hat{Y})$ is $2^{-\Omega(k^{\gamma/8})}$-close to a uniform string on ℓ bits. Thus, by Theorem 5.1, equation (8.2) holds except with probability $(2^{-\ell} + 2^{-\Omega(k^{\gamma/8})}) \cdot \mathrm{poly}(n) = 2^{-\Omega(\ell)}$, where we used our assumption on ℓ.

We now proceed to the induction step. Let $0 \leq i < i_{\mathsf{bot}}(T_X) - 1$. Let $X' \subset X_{\mathsf{FI}}$, $Y' \subset Y_{\mathsf{FI}}$ be deficiency $i \cdot t$ subsources. We want to show that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X',Y')}[\forall j \in \{i,\ldots,i_{\mathsf{bot}}(T_X)-1\} \quad v_j(T_X) \text{ favors its good son}] \geq 1-\varepsilon_i.$$

By the induction hypothesis, for any deficiency $(i+1)t$ subsources $X'' \subset X_{\mathsf{FI}}$, $Y'' \subset Y_{\mathsf{FI}}$, it holds that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X'',Y'')}[\forall j \in \{i+1,\ldots,i_{\mathsf{bot}}(T_X)-1\} \quad v_j(T_X) \text{ favors its good son}] \geq 1-\varepsilon_{i+1}.$$

As was done in the basis of the induction, we consider two cases. If $\mathsf{label}(\mathsf{leftSon}(v_i(T_X))) = \mathsf{F}$, then by Claim 8.3

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X_{\mathsf{FI}},Y_{\mathsf{FI}})}[v_i(T_X) \ (x,y)\text{-favors its good son}] = 1.$$

Since $X' \subset X_{\mathsf{FI}}$ and $Y' \subset Y_{\mathsf{FI}}$, the same holds for $(x,y) \sim (X',Y')$. Thus, by the induction hypothesis

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(X',Y')}[\forall j \in \{i,\ldots,i_{\mathsf{bot}}(T_X)-1\} \quad v_j(T_X) \text{ favors its good son}] \geq 1-\varepsilon_{i+1} \geq 1-\varepsilon_i.$$

Consider now the case $\mathsf{label}(\mathsf{leftSon}(v_i(T_X))) \neq \mathsf{F}$. By the observation made at the beginning of the proof, in this case the good son of $v_i(T_X)$ is its left son. Thus, $v_i(T_X)$ (x,y)-favors its good son if and only if

$$\mathsf{Response}\left(x_{v_i(T_X)}, y, \mathsf{GoodSonCh}\left(x_{\mathsf{leftSon}(v_i(T_X))}, y\right)\right) = \mathsf{hasEntropy}. \tag{8.3}$$

By Theorem 5.1, it is enough to show that for any deficiency t subsources $\hat{X} \subset X'$, $\hat{Y} \subset Y'$, it holds that $\mathsf{GoodSonCh}(\hat{X}_{\mathsf{leftSon}(v_i(T_X))}, \hat{Y})$ is close to having min-entropy ℓ. Since $\hat{X}$ is a deficiency t subsource of X', and since X' is a deficiency $i \cdot t$ subsource of X_{FI}, we have that $\hat{X}$ is a deficiency $(i+1)t$ subsource of X_{FI}. Similarly, $\hat{Y}$ is a deficiency $(i+1)t$ subsource of Y_{FI}. Thus, by the induction hypothesis,

$$\mathop{\mathbf{Pr}}_{(x,y)\sim(\hat{X},\hat{Y})}[\forall j \in \{i+1,\ldots,i_{\mathsf{bot}}(T_X)-1\} \quad v_j(T_X) \text{ favors its good son}] \geq 1-\varepsilon_{i+1}.$$

By the above equation and by the definition of $\mathsf{GoodSonCh}$, except for probability ε_{i+1} over $(x,y) \sim (\hat{X},\hat{Y})$, it holds that $\mathsf{BExt}(x_{v_{\mathsf{bot}}(T_X)}, y)$ appears as a row in $\mathsf{GoodSonCh}(x_{v_{i+1}(T_X)}, y)$.

Since $\hat{X}$ is a deficiency $O((i+1)t + \ell \log^2 n) = O(\ell \log^2 n)$ subsource of X and since $X_{v_{\mathsf{bot}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/8})$-block-source, $\hat{X}_{v_{\mathsf{bot}}(T_X)}$ is $2^{-\Omega(k^{1/8})}$-close to an $\Omega(k^{1/8})$-block-source. Further, since $\hat{Y}$ is a deficiency $O(\ell \log^2 n)$ subsource of Y and since $\mathbf{H}_\infty(Y) \geq k$, $\mathbf{H}_\infty(\hat{Y}) = \Omega(k)$. Theorem 4.1 and the remark following it then imply that $\mathsf{BExt}(\hat{X}_{v_{\mathsf{bot}}(T_X)}, \hat{Y})$ is $2^{-\Omega(k^{\gamma/8})}$-close to a uniform string on ℓ bits. Thus, $\mathsf{GoodSonCh}(\hat{X}_{\mathsf{leftSon}(v_i(T_X))}, \hat{Y})$ is $(\varepsilon_{i+1} + 2^{-\Omega(k^{\gamma/8})})$-close to having min-entropy ℓ. Therefore, by Theorem 5.1, equation (8.3) holds except with probability

$$\left(2^{-\ell} + \varepsilon_{i+1} + 2^{-\Omega(k^{\gamma/8})}\right) \cdot \mathrm{poly}(n) = \left(2^{-\Omega(\ell)} + \varepsilon_{i+1}\right) \cdot \mathrm{poly}(n) = \varepsilon_i.$$

This concludes the proof of the claim. ∎

8.2. Analysis of Step 2. Informally speaking, in this section we prove that the subextractor correctly identifies $v_{\mathsf{mid}}(T_X)$ in some carefully chosen subsources of $X_{\mathsf{Fl}}, Y_{\mathsf{Fl}}$. More precisely, we would have wanted to prove a statement of the following form.

A wishful claim. There exist low-deficiency subsources $X' \subset X_{\mathsf{Fl}}$, $Y' \subset Y_{\mathsf{Fl}}$ such that with high probability over $(x,y) \sim (X', Y')$, $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y) = v_{\mathsf{mid}}(T_X)$.

Unfortunately, we will not be able to prove this statement. Nevertheless, we will be able to prove the same statement for X', Y' that have *high-deficiency* in $X_{\mathsf{Fl}}, Y_{\mathsf{Fl}}$. Still, X', Y' will have enough entropy and structure to carry out the rest of the analysis. Furthermore, the error term that we are carrying will not cause any harm even after moving to these high-deficiency subsources.

For $\alpha \in \mathsf{supp}((X_{\mathsf{Fl}})_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))})$ and $\beta \in \mathsf{supp}((Y_{\mathsf{Fl}})_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))})$, we define

$$X_\alpha = X_{\mathsf{Fl}} \mid ((X_{\mathsf{Fl}})_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))} = \alpha),$$
$$Y_\beta = Y_{\mathsf{Fl}} \mid ((Y_{\mathsf{Fl}})_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))} = \beta).$$

Let B be the set of all $(x,y) \in \mathsf{supp}((X_{\mathsf{Fl}}, Y_{\mathsf{Fl}}))$ such that

$$\begin{aligned} &\exists i \in \{0,\ldots,i_{\mathsf{bot}}(T_X)\} \quad v_i(x,y) \neq v_i(T_X) \quad \vee \\ &\exists i \in \{0,\ldots,i_{\mathsf{bot}}(T_Y)\} \quad u_i(x,y) \neq u_i(T_Y). \end{aligned} \tag{8.4}$$

By Claim 8.4,

$$\mathbf{Pr}[(X_{\mathsf{Fl}}, Y_{\mathsf{Fl}}) \in B] \leq 2^{-\Omega(\ell)}.$$

Thus, by averaging, there exist α, β such that

$$\mathbf{Pr}[(X_\alpha, Y_\beta) \in B] \leq 2^{-\Omega(\ell)}.$$

These are the subsources $X_\alpha \subset X_{\mathsf{Fl}}$, $Y_\beta \subset Y_{\mathsf{Fl}}$ that we will work with. We think of $(x,y) \in B$ as an "error" and ignore this event for now. We later accumulate the error coming from this event while making sure to treat the error correctly when moving into subsources of (X_α, Y_β). More precisely, note that by moving to a deficiency d subsource, an error of ε in the source can "grow" to at most $2^d \cdot \varepsilon$ restricted to the subsource. Since the error term is $2^{-c\ell}$, for some constant c, and since we will move to deficiency $c'\ell$ subsources, where $c' < c$ is another constant, the error will remain $2^{(c'-c)\ell} = 2^{-\Omega(\ell)}$ in the subsources that we will restrict to. To summarize, we assume that (8.4) holds. In particular, we assume that $v_{i_{\mathsf{top}}(T_X)}(X_\alpha, Y_\beta) = v_{\mathsf{top}}(T_X)$, $v_{i_{\mathsf{mid}}(T_X)}(X_\alpha, Y_\beta) = v_{\mathsf{mid}}(T_X)$, etc.

Recall that $v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)$ is defined to be the node v in $p_{\mathsf{observed}}(x,y)$, with the largest depth, for which

$$\mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_v, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{hasEntropy}. \tag{8.5}$$

If no such node v exists, v is defined to be $\mathsf{root}(T_X)$. Thus, to show that $v_{\mathsf{mid}}(T_X)$ is correctly identified on low-deficiency subsources of X_α, Y_β, we first show that there exist low-deficiency subsources $X_{\alpha,\beta} \subset X_\alpha$, $Y_{\alpha,\beta} \subset Y_\beta$ such that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, equation (8.5) does not hold with $v = v_i(x,y)$ for all $i > i_{\mathsf{mid}}(T_X)$. This is the content of the following claim. Afterward, in Claim 8.8, we show that with high probability over $(x,y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, equation (8.5) holds with $v = v_{i_{\mathsf{mid}}(T_X)}(x,y) = v_{\mathsf{mid}}(T_X)$.

CLAIM 8.5. *There exist deficiency $O(\ell' \log^2 n)$ subsources $X_{\alpha,\beta} \subset X_\alpha$, $Y_{\alpha,\beta} \subset Y_\beta$ such that with probability $1 - 2^{-\Omega(\ell)}$ over $(x, y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$, it holds that*

$$\forall i > i_{\mathsf{mid}}(T_X) \quad \mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_{v_i(x,y)}, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{fixed}. \tag{8.6}$$

Toward proving Claim 8.5, we start by proving the following two claims.

CLAIM 8.6. *There exists a deficiency $\log n$ subsource $X'_\alpha \subset X_\alpha$ such that $q_{\mathsf{observed}}(X'_\alpha, Y_\beta)$ is fixed to a constant.*

Proof. Recall that $q_{\mathsf{observed}}(X_\alpha, Y_\beta)$ is the (random) path

$$u_0(X_\alpha, Y_\beta), \ldots, u_{i_{\mathsf{bot}}(T_Y)}(X_\alpha, Y_\beta), u_{i_{\mathsf{bot}}(T_Y)+1}(X_\alpha, Y_\beta), \ldots, u_{\log(n)-1}(X_\alpha, Y_\beta).$$

By (8.4), for all $0 \le i \le i_{\mathsf{bot}}(T_Y)$ it holds that $u_i(X_\alpha, Y_\beta) = u_i(T_Y)$. In particular, for any such i, $u_i(X_\alpha, Y_\beta)$ is fixed to a constant. We now consider indices $i > i_{\mathsf{bot}}(T_Y)$. Consider first $i = i_{\mathsf{bot}}(T_Y) + 1$. In this case, $u_i(X_\alpha, Y_\beta)$ is one of the two sons of $u_{i_{\mathsf{bot}}(T_Y)}(X_\alpha, Y_\beta) = u_{\mathsf{bot}}(T_Y)$. Recall that the decision regarding which son will be on $q_{\mathsf{observed}}(X_\alpha, Y_\beta)$ is based on whether or not

$$\mathsf{Response}\left((Y_\beta)_{u_{\mathsf{bot}}(T_Y)}, X_\alpha, \mathsf{GoodSonCh}\left((Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))}, X_\alpha\right)\right) = \mathsf{hasEntropy}. \tag{8.7}$$

Since $u_{\mathsf{bot}}(T_Y)$ and $\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))$ are descendants of $\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))$, as follows by the definition of entropy-trees, and since $(Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{mid}}(T_Y))}$ is fixed to β, it holds that $(Y_\beta)_{u_{\mathsf{bot}}(T_Y)}$ and $(Y_\beta)_{\mathsf{leftSon}(u_{\mathsf{bot}}(T_Y))}$ are fixed to constants. Thus, (8.7) is determined only by X_α. As (8.7) reveals one bit of information on X_α, by Fact 4.4, there exists a deficiency 1 subsource $X' \subset X_\alpha$ such that $u_i(X', Y_\beta)$ is fixed to a constant.

We now repeat this argument for $i = i_{\mathsf{bot}}(T_Y) + 2, \ldots, \log(n) - 1$. Each time we make sure that the next descendant of $u_{\mathsf{bot}}(T_Y)$ is fixed to a constant on a low-deficiency subsource of X_α with Y_β. Since we repeat this process at most $\log n$ times, we will eventually obtain a deficiency $\log n$ subsource $X'_\alpha \subset X_\alpha$ such that $q_{\mathsf{observed}}(X'_\alpha, Y_\beta)$ is fixed to a constant, as desired. Accounting for the error, note that since $\ell = \omega(\log n)$, it holds that

$$\mathbf{Pr}\left[(X'_\alpha, Y_\beta) \in B\right] \le 2^{-\Omega(\ell)}. \tag{8.8}$$

∎

CLAIM 8.7. *For any $i > i_{\mathsf{mid}}(T_X)$,*

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}, (Y_\beta)_{q_{\mathsf{observed}}(X'_\alpha, Y_\beta)}\right)$$

is a deterministic function of Y_β.

Proof. By Claim 8.6, we have that $q_{\mathsf{observed}}(X'_\alpha, Y_\beta)$ is fixed to a constant. Thus, it suffices to show that $(X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}$ is a deterministic function of Y_β. We start by considering $i = i_{\mathsf{mid}}(T_X) + 1$. In this case, $v_i(X'_\alpha, Y_\beta)$ is fixed to a constant. Indeed, since $i_{\mathsf{mid}}(T_X) + 1 \le i_{\mathsf{bot}}(T_X)$, it holds by (8.4) that $v_i(X_\alpha, Y_\beta) = v_i(T_X)$ and so, since X'_α is a subsource of X_α, $v_i(X'_\alpha, Y_\beta) = v_i(T_X)$.

The case $i > i_{\mathsf{mid}}(T_X) + 1$ follows by a different logic, similar to that used in the proof of Claim 8.6. Let us first consider $i = i_{\mathsf{mid}}(T_X) + 2$. Recall that $(X_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ is fixed to α. Thus, also $(X'_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ is fixed to α. Now, $v_i(X'_\alpha, Y_\beta)$ is defined to be one of the two sons of $\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))$ according to the Boolean value of the expression

$$\mathsf{Response}\left((X'_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}, Y_\beta, \mathsf{GoodSonCh}\left((X'_\alpha)_{\mathsf{leftSon}(\mathsf{leftSon}(v_{\mathsf{mid}}(T_X)))}, Y_\beta\right)\right) = \mathsf{hasEntropy}.$$

Since $(X'_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ is fixed to a constant, the above equation is determined only by Y_β. This shows that $v_i(X'_\alpha, Y_\beta)$ is a deterministic function of Y_β for $i = i_{\mathsf{mid}}(T_X) + 2$. One can now use a similar argument to show that the same holds for any $i > i_{\mathsf{mid}}(T_X) + 1$.

To conclude the proof of the claim, we need to show that $(X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}$ is a deterministic function of Y_β for $i > i_{\mathsf{mid}}(T_X)$. By the above, for any such i, $v_i(X'_\alpha, Y_\beta)$ is a descendant of $\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))$ determined only by Y_β. The claim then follows as $(X'_\alpha)_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))}$ is fixed to a constant. ∎

We are now ready to prove Claim 8.5.

Proof of Claim 8.5. By Claim 8.7,

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha, Y_\beta)}, (Y_\beta)_{q_{\mathsf{observed}}(X'_\alpha, Y_\beta)}\right)$$

is a deterministic function of Y_β for all $i > i_{\mathsf{mid}}(T_X)$. Thus, there exists a deficiency $\ell' \log^2 n$ subsource $Y'_\beta \subset Y_\beta$ such that for all $i > i_{\mathsf{mid}}(T_X)$,

$$\mathsf{NodePathCh}\left((X'_\alpha)_{v_i(X'_\alpha, Y'_\beta)}, (Y'_\beta)_{q_{\mathsf{observed}}(X'_\alpha, Y'_\beta)}\right)$$

is fixed to a constant. Recall that $\ell' \log^2 n = \ell/c_2$. We set the constant c_2 to be smaller than the constant hidden in the Ω-notation in the error term of (8.8) so that $\mathbf{Pr}[(X'_\alpha, Y'_\beta) \in B] \le 2^{-\Omega(\ell)}$. By Theorem 5.1, there exist deficiency $2\ell' \log^2 n + 2$ subsources $X_{\alpha,\beta} \subset X'_\alpha$, $Y_{\alpha,\beta} \subset Y'_\beta$, such that with probability $1 - 2^{-\Omega(\ell)} \cdot 2^{4\ell' \log^2 n + 4} = 1 - 2^{\Omega(\ell)}$ over $(x, y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$ equation (8.6) holds, where the last equation holds for an appropriate choice of the constant c_2.

We note that this application of Theorem 5.1 is valid as both $X_{\alpha,\beta}$, $Y_{\alpha,\beta}$ are 1/4-close to having min-entropy $\Omega(\log^{10} n)$. Indeed, $X_{\alpha,\beta}$ is a deficiency $O(\ell' \log^2 n)$ subsource of $X_\alpha = X_{\mathsf{FI}} \mid ((X_{\mathsf{FI}})_{\mathsf{leftSon}(v_{\mathsf{mid}}(T_X))} = \alpha)$. Since X_{FI} is a deficiency $O(\ell \log^2 n)$ subsource of X and since $X_{v_{\mathsf{mid}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source, our choice of k implies that $(X_{\mathsf{FI}})_{v_{\mathsf{mid}}(T_X)}$ is also $\Omega(k^{1/4})$-close to an $\Omega(k^{1/4})$-block-source. Thus, X_α is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$. Therefore, $X_{\alpha,\beta}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(\log^{10} n)$. A similar argument can be used to show that $Y_{\alpha,\beta}$ is 1/4-close to having min-entropy $\Omega(\log^{10} n)$, as required by Theorem 5.1. ∎

Claim 8.8. *With probability* $1 - O(\varepsilon)$ *over* $(x, y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$ *it holds that*

$$\mathsf{Response}\left(x, y, \mathsf{NodePathCh}\left(x_{v_{i_{\mathsf{mid}}(T_X)}(x,y)}, y_{q_{\mathsf{observed}}(x,y)}\right)\right) = \mathsf{hasEntropy}. \tag{8.9}$$

Proof. Let B be the event defined in (8.4). As usual, we consider $(x, y) \in B$ an "error" and ignore it for now. In particular, $v_{i_{\mathsf{mid}}(T_X)}(x, y) = v_{\mathsf{mid}}(T_X)$ and the path $q_{\mathsf{observed}}(x, y)$ is assumed to contain $u_{\mathsf{top}}(T_Y)$. Thus, $\mathsf{NodePathCh}(x_{v_{i_{\mathsf{mid}}(T_X)}(x,y)}, y_{q_{\mathsf{observed}}(x,y)})$ contains $\mathsf{BExt}\left(y_{u_{\mathsf{top}}(T_Y)}, x_{v_{\mathsf{mid}}(T_X)}\right)$ as a row.

Recall that $(Y)_{u_{\mathsf{top}}(T_Y)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(\sqrt{k})$-block-source. As Y_{FI} is a deficiency $O(\ell \log^2 n)$ subsource of Y, our assumption on k implies that $(Y_{\mathsf{FI}})_{u_{\mathsf{top}}(T_Y)}$ is also $2^{-\Omega(k^{1/4})}$-close to an $\Omega(\sqrt{k})$-block-source. By a similar argument, $(Y_{\mathsf{FI}})_{u_{\mathsf{mid}}(T_Y)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source. Thus, $(Y_\beta)_{u_{\mathsf{mid}}(T_Y)}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$, and so $(Y_\beta)_{u_{\mathsf{top}}(T_Y)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source. As $Y_{\alpha,\beta}$ is a deficiency $O(\ell' \log^2 n)$ subsource of Y_β, our choice of ℓ' implies that $(Y_{\alpha,\beta})_{u_{\mathsf{top}}(T_Y)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source.

Recall that $X_{v_{\mathsf{mid}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source. Since X_{FI} is a deficiency $O(\ell \log^2 n)$ subsource of X, $(X_{\mathsf{FI}})_{v_{\mathsf{mid}}(T_X)}$ is also $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source. Thus, $(X_\alpha)_{v_{\mathsf{mid}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$. Our choice of ℓ' then implies that $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$.

Let $\hat{X} \subset X_{\alpha,\beta}$, $\hat{Y} \subset Y_{\alpha,\beta}$ be any deficiency $20\ell' \log n$ subsources. By our assumption on k, we have that $\hat{Y}_{u_{\mathsf{top}}(T_Y)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source and that $\hat{X}_{v_{\mathsf{mid}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$. Thus, $\mathsf{BExt}(\hat{Y}_{u_{\mathsf{top}}(T_Y)}, \hat{X}_{v_{\mathsf{mid}}(T_X)})$ is $2^{-(k^{\gamma/4})}$-close to a uniform string on ℓ' bits. Thus, continuing to ignore the error set B, $\mathsf{NodePathCh}(\hat{X}_{v_{i_{\mathsf{mid}}(T_X)}(\hat{X},\hat{Y})}, \hat{Y}_{q_{\mathsf{observed}}(\hat{X},\hat{Y})})$ is $2^{-(k^{\gamma/4})}$-close to having min-entropy ℓ'. Theorem 5.1 then implies that (8.9) holds except with probability

$$\left(2^{-\ell'} + 2^{-\Omega(k^{\gamma/4})}\right) \cdot \mathrm{poly}(n) = O(\varepsilon).$$

The error term coming from the set B contributes additional $O(\varepsilon)$ to the total error. This concludes the proof of the claim. ∎

8.3. Analysis of Step 3. Recall that the output of the subextractor is defined as

$$\mathsf{SubExt}(x, y) = \mathsf{BExt}\left(x_{v_{\mathsf{mid}}^{\mathsf{observed}}(x,y)} \circ x, y\right).$$

By Claims 8.5 and 8.8, we have that except with probability $O(\varepsilon)$ over $(x, y) \sim (X_{\alpha,\beta}, Y_{\alpha,\beta})$ it holds that $v_{\mathsf{mid}}^{\mathsf{observed}}(x, y) = v_{\mathsf{mid}}(T_X)$. Recall that $v_{\mathsf{mid}}(T_X)$ is a descendant of $\mathsf{leftSon}(v_{\mathsf{top}}(T_X))$. Further, recall that $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(k^{1/4})$. As $X_{v_{\mathsf{top}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(\sqrt{k})$-block-source, we have that $(X_{\alpha,\beta})_{v_{\mathsf{top}}(T_X)}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source. In particular, this implies that $(X_{\alpha,\beta})_{v_{\mathsf{mid}}(T_X)} \circ X_{\alpha,\beta}$ is $2^{-\Omega(k^{1/4})}$-close to an $\Omega(k^{1/4})$-block-source.

As $Y_{\alpha,\beta}$ is $2^{-\Omega(k^{1/4})}$-close to having min-entropy $\Omega(\sqrt{k})$, Theorem 4.1 implies that

$$\mathsf{SubExt}(X_{\alpha,\beta}, Y_{\alpha,\beta}) = \mathsf{BExt}\left((X_{\alpha,\beta})_{v_{\mathsf{mid}}^{\mathsf{observed}}(X_{\alpha,\beta},Y_{\alpha,\beta})} \circ X_{\alpha,\beta}, Y_{\alpha,\beta}\right)$$

is $(2^{-\Omega(k^{\gamma/4})} + O(\varepsilon))$-close to uniform. The proof of the claim then follows as $2^{-\Omega(k^{\gamma/4})} = O(\varepsilon)$. Note further that by Theorem 4.1, the output length of SubExt is $\Omega(k^{1/4})$. This proves Theorem 8.1.

Acknowledgments. I wish to thank Ran Raz and Avi Wigderson for their warm encouragement. On a personal note, it is uncustomary to acknowledge one's partner in life in mathematical papers. However, given that this paper was intensively written in the last month of my wife's pregnancy and in the first month of parenthood to the newborn baby girl Meshi and to our sweet Yahli, I will allow myself to make an exception—thank you, Orit! Your support and belief in my abilities are uncanny.

REFERENCES

[1] H. L. Abbott, *Lower bounds for some Ramsey numbers*, Discrete Math., 2 (1972), pp. 289–293.
[2] N. Alon, *The Shannon capacity of a union*, Combinatorica, 18 (1998), pp. 301–310.
[3] B. Barak, *A Simple Explicit Construction of an $n^{\tilde{O}(\log n)}$-Ramsey Graph*, preprint, https://arxiv.org/abs/math/0601651, 2006.

[4] B. Barak, R. Impagliazzo, and A. Wigderson, *Extracting randomness using few independent sources*, SIAM J. Comput., 36 (2006), pp. 1095–1118, https://doi.org/10.1137/S0097539705447141.
[5] B. Barak, G. Kindler, R. Shaltiel, B. Sudakov, and A. Wigderson, *Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors*, J. ACM, 57 (2010), 20.
[6] B. Barak, A. Rao, R. Shaltiel, and A. Wigderson, 2-*source dispersers for* $n^{o(1)}$ *entropy, and Ramsey graphs beating the Frankl-Wilson construction*, Ann. of Math. (2), 176 (2012), pp. 1483–1543.
[7] A. Ben-Aroya, D. Doron, and A. Ta-Shma, *Explicit two-source extractors for near-logarithmic min-entropy*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2016, 088.
[8] E. Ben-Sasson and N. Zewi, *From affine to two-source extractors via approximate duality*, in Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, ACM, New York, 2011, pp. 177–186.
[9] J. Bourgain, *More on the sum-product phenomenon in prime fields and its applications*, Int. J. Number Theory, 1 (2005), pp. 1–32.
[10] E. Chattopadhyay and X. Li, *Explicit non-malleable extractors, multi-source extractors, and almost optimal privacy amplification protocols*, in Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE, Washington, DC, 2016, pp. 158–167.
[11] E. Chattopadhyay and D. Zuckerman, *Explicit two-source extractors and resilient functions*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2015, p. 119.
[12] B. Chor and O. Goldreich, *Unbiased bits from sources of weak randomness and probabilistic communication complexity*, SIAM J. Comput., 17 (1988), pp. 230–261, https://doi.org/10.1137/0217015.
[13] F. Chung, *A note on constructive methods for Ramsey numbers*, J. Graph Theory, 5 (1981), pp. 109–113.
[14] G. Cohen, *Local correlation breakers and applications to three-source extractors and mergers*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2015, 038.
[15] G. Cohen, *Making the most of advice: New correlation breakers and their applications*, in Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE, Washington, DC, 2016, pp. 188–196.
[16] G. Cohen, *Two-source extractors for quasi-logarithmic min-entropy and improved privacy amplification protocols*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2016, 114.
[17] G. Cohen and L. Schulman, *Extractors for near logarithmic min-entropy*, in Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE, Washington, DC, 2016, pp. 178–187.
[18] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc., 53 (1947), pp. 292–294.
[19] A. Fiat and M. Naor, *Implicit* $O(1)$ *probe search*, SIAM J. Comput., 22 (1993), pp. 1–10, https://doi.org/10.1137/0222001.
[20] P. Frankl, *A constructive lower bound for some Ramsey numbers*, Ars Combin., 3 (1977), pp. 297–302.
[21] P. Frankl and R. M. Wilson, *Intersection theorems with geometric consequences*, Combinatorica, 1 (1981), pp. 357–368.
[22] A. Gabizon and R. Shaltiel, *Increasing the output length of zero-error dispersers*, in Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, Springer, New York, 2008, pp. 430–443.
[23] P. Gopalan, *Constructing Ramsey graphs from Boolean function representations*, Combinatorica, 34 (2014), pp. 173–206.
[24] R. Gradwohl, G. Kindler, O. Reingold, and A. Ta-Shma, *On the error parameter of dispersers*, in Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, Springer, New York, 2005, pp. 294–305.
[25] V. Grolmusz, *Low rank co-diagonal matrices and Ramsey graphs*, J. Combin., 7 (2001), pp. R15–R15.
[26] Y. Kalai, X. Li, A. Rao, and D. Zuckerman, *Network extractor protocols*, in Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE, Washington, DC, 2008, pp. 654–663.

[27] X. Li, *Improved constructions of three source extractors*, in Proceedings of the 26th IEEE Annual Conference on Computational Complexity, IEEE, Washington, DC, 2011, pp. 126–136.
[28] X. Li, *Extractors for a constant number of independent sources with polylogarithmic min-entropy*, in Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science, IEEE, Washington, DC, 2013, pp. 100–109.
[29] X. Li, *Improved constructions of two-source extractors*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2015, p. 125.
[30] X. Li, *Three-source extractors for polylogarithmic min-entropy*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2015, p. 190.
[31] X. Li, *Improved non-malleable extractors, non-malleable codes and independent source extractors*, in Electronic Colloquium on Computational Complexity (ECCC), Weizmann Institute of Science, Rehovot, Israel, 2016, 115.
[32] Z. Nagy, *A constructive estimation of the Ramsey numbers*, Mat. Lapok, 23 (1975), pp. 301–302.
[33] M. Naor, *Constructing Ramsey Graphs from Small Probability Spaces*, IBM Research Report RJ 8810, IBM Thomas J. Watson Research Division, 1992.
[34] P. Pudlák and V. Rödl, *Pseudorandom sets and explicit constructions of Ramsey graphs*, Quad. Mat, 13 (2004), pp. 327–346.
[35] F. P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. (2), 30 (1929), pp. 264–286.
[36] A. Rao, *A 2-source almost-extractor for linear entropy*, in Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, Springer, New York, 2008, pp. 549–556.
[37] A. Rao, *Extractors for a constant number of polynomially small min-entropy independent sources*, SIAM J. Comput., 39 (2009), pp. 168–194, https://doi.org/10.1137/060671218.
[38] R. Raz, *Extractors with weak random seeds*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing, ACM, New York, 2005, pp. 11–20.
[39] A. Vitanov, F. Dupuis, M. Tomamichel, and R. Renner, *Chain rules for smooth min-and max-entropies*, IEEE Trans. Inform. Theory, 59 (2013), pp. 2603–2612.