Electronic Colloquium on Computational Complexity, Report No. 178 (2015)

Extractors for Sumset Sources

Eshan Chattopadhyay*
Department of Computer Science,
University of Texas at Austin
eshanc@cs.utexas.edu

Xin Li
Department of Computer Science,
John Hopkins University
lixints@cs.jhu.edu

November 10, 2015

Abstract

We propose a new model of weak random sources which we call *sumset sources*. A sumset source $\mathbf{X}$ is the sum of C independent sources $\mathbf{X}_1, \ldots, \mathbf{X}_C$, where each $\mathbf{X}_i$ is an n-bit source with min-entropy k. We show that extractors for this class of sources can be used to give extractors for most classes of weak sources that have been studied previously, including independent sources, affine sources (which generalizes oblivious bit-fixing sources), small space sources, total entropy independent sources, and interleaved sources. This provides a unified approach for randomness extraction.

A known extractor for this class of sources, prior to our work, is the Paley graph function introduced by Chor and Goldreich [CG88], which works for the sum of 2 independent sources, where one has min-entropy at least $0.51n$ and the other has min-entropy $O(\log n)$. To the best of our knowledge, the only other known construction is from the work of Kamp et al. [KRVZ11], which can extract from the sum of exponentially many independent sources.

Our main result is an explicit extractor for the sum of C independent sources for some large enough constant C, where each source has min-entropy polylog(n). We then use this extractor to obtain the following results for other well studied classes of sources.

- **Small-space sources:** This is the class of sources generated by a small width branching program. Previously the best known extractor by Kamp et al. [KRVZ11] requires min-entropy $k \geq n^{1-\delta}$ even for space 1, where $\delta > 0$ is a small constant. We improve the min-entropy to $k \geq 2^{\log^{0.51}(n)} s^{1.1}$ for space s, which is $n^{o(1)}$ for $s = n^{o(1)}$.
- **Affine Sources:** This constitutes the class of sources that are uniform over some affine subspace in $\mathbb{F}_2^n$. A direct corollary of our sumset extractor gives an explicit affine extractor for entropy polylog(n), matching the recent work of Li [Li15a].
- **Interleaved Sources:** We obtain new results on extracting from an unknown interleaving of the bits of C independent sources. This model was studied by Raz and Yehudayoff [RY11] in the context of proving circuit lower bounds, and subsequently by Chattopadhyay and Zuckerman [CZ15b]. Previous results require at least one source to have min-entropy $(1-\delta)n$ for a small constant $\delta > 0$. We give explicit extractors for the interleaving of a constant number of sources each with polylogarithmic min-entropy.

We also give improved extractors for total entropy independent sources, introduced in [KRVZ11], and a simple extractor for somewhere-2 sources, which generalizes the model of 2-independent sources to a large collection of independent sources with the guarantee that at least two sources contain polylogarithmic min-entropy.

*Partially supported by NSF Grant CCF-1526952.

ISSN 1433-8092

1 Introduction

The use of randomness is widespread in various branches of computer science, such as algorithms, data structures, distributed computing, cryptography and many more. Most of these applications in fact require the random bits to be uniform and uncorrelated. However, natural random sources are often biased and only contain some small amount of entropy, and in cryptographic applications even original uniform random sources can leak information to an adversary as a result of side channel attacks. This motivates the broad area of randomness extraction, which deals with the important problem of designing efficient algorithms (known as randomness extractors) that can extract almost uniform random bits from defective random sources.

To formally define defective or weak random sources, we use the following standard model.

Definition 1.1. *The min-entropy of a source* $\mathbf{X}$ *is defined to be:* $H_\infty(\mathbf{X}) = min_x\{-\log(\Pr[\mathbf{X} = x])\}$. *The min-entropy rate of a source* $\mathbf{X}$ *on* $\{0,1\}^n$ *is defined to be* $H_\infty(\mathbf{X})/n$. *Any source* $\mathbf{X}$ *on* $\{0,1\}^n$ *with min-entropy at least* k *is called an* (n,k)*-source.*

We use the standard statistical distance to measure the distance between two distributions.

Definition 1.2. *The statistical distance between two distributions* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *over some universal set* Ω *is defined as* $|\mathcal{D}_1 - \mathcal{D}_2| = \frac{1}{2}\sum_{x\in\Omega} |\mathbf{Pr}[\mathcal{D}_1 = x] - \mathbf{Pr}[\mathcal{D}_2 = x]|$. *We say* $\mathcal{D}_1$ *is* ϵ*-close to* $\mathcal{D}_2$ *if* $|\mathcal{D}_1 - \mathcal{D}_2| \leq \epsilon$ *and denote it by* $\mathcal{D}_1 \approx_\epsilon \mathcal{D}_2$.

We are now ready to define extractors for a class of sources.

Definition 1.3. *We say that an efficiently computable function* $f : \{0,1\}^n \to \{0,1\}^m$ *is an extractor for a class of sources* $\mathcal{X}$ *with error* ϵ*, if for any source* $\mathbf{X} \in \mathcal{X}$, $|f(\mathbf{X}) - \mathbf{U}_m| \leq \epsilon$. *Here* U_m *denotes the uniform distribution on* $\{0,1\}^m$.

Given the above definition, a natural goal would be to design a deterministic extractor for the class of (n,k)-sources. However, a simple argument shows that such an extractor cannot exist even for $k = n - 1$. This has motivated researchers to consider two different approaches. The first approach is to allow the extractor to have a small uniform independent random seed, and these extractors are called *seeded extractors*. Seeded extractors were first defined by Nisan and Zuckerman [NZ96], and through a long line of research we now have almost optimal constructions (e.g., [GUV09]). The second approach, which is the focus of this paper, is to design *seedless extractors* for more restricted sources. Here, the goal is to identify the most general class of sources that allows the construction of explicit deterministic randomness extractors. This question can be dated back to von Neumann [vN51], and continued interest in this question over the past three decades has led to many fascinating new techniques and results.

One important class of sources that has received a lot of attention is the class of multiple independent sources. Such extractors are particularly useful for distributed computing and cryptographic applications which involve multiple parties [KLRZ08, KLR09]. Here, the probabilistic method can be used to show the existence of an extractor for 2 independent sources, each with min-entropy $\log n + O(1)$. However to come up with an explicit extractor that matches this bound is extremely challenging. By a long line of successful research [CG88, BIW06, BKS+10, Bou05, Raz05, Rao09a, BRSW12, Li13b, Li13a, Li15c, CZ15a, Li15b], we now have extractors for 2 independent sources with each source containing min-entropy $\log^C n$ for some constant $C > 1$.

Many other interesting models have been investigated, and we briefly mention some examples here. One well studied class of sources is known as bit fixing sources [CGH+85, KZ07, GRS06,

Rao09b], which are sources that are obtained by fixing some unknown bits of a uniform random string. Explicit extractors for such sources have found applications in exposure-resilient cryptography [CGH+85, KZ07]. Generalizing these sources, another well studied class of sources is affine sources [GR08, Bou07, Rao09b, Yeh11, BK12, Sha11, Li11, Li15a], which are the uniform distribution over some unknown affine subspace of a vector space. Extractors for affine sources are shown to be related to two-source extractors [BSZ11], and imply the best known lower bounds for general Boolean circuits [DK11, FGHK15]. Other classes of sources that have been previously studied include samplable sources [TV00, Vio14], which are sources that are generated by small circuits or efficient algorithms; interleaved sources [RY11, CZ15b], which are a generalization of independent sources where the bits of the sources are mixed in some arbitrary order; and small-space sources [KRVZ11], where the sources are generated by a small width branching program.

In this paper we propose a new model of weak sources which we call *sumset sources*. Informally, this is the class of sources which are the sum (XOR) of independent sources. We show that most of the classes of sources studied before (as we mentioned above), are either special cases in our new model, or can be reduced to our new model. This further reduces the assumptions made on weak sources, and provides a unified framework for designing extractors for well studied classes of sources. We then construct explicit extractors for sumset sources and apply them to other classes of sources studied before. In several cases we obtain substantial improvements over previous constructions. We now formally define sumset sources.

1.1 Sumset Sources

Definition 1.4. *For any two strings $x, y \in \{0,1\}^n$, define $x + y$ to be the bit wise* XOR *of the two strings.*

Definition 1.5 ((n, k, C)-sumset source)**.** *A weak source $\mathbf{X}$ is called an (n, k, C)-sumset source if $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_C$, where $\mathbf{X}_1, \ldots, \mathbf{X}_C$ are independent (n, k)-sources.*

A well known extractor for this class of sources is based on the Paley graph function introduced by Chor and Goldreich [CG88] and works for the sum of 2 independent sources, with one having min-entropy at least $> n/2$ and the other having min-entropy $> \log n$. On the other extreme, the work of Kamp et al. [KRVZ11] shows how to extract when $\mathbf{X}$ is a sum of exponentially many sources ($C = 2^{O(n)}$). To the best of our knowledge, there is no other known explicit construction for $2 \leq C \leq 2^{O(n)}$. Further, it is not clear if one can use the probabilistic method to prove the existence of such extractors.

Our main result is an explicit construction of an extractor for the sum of a constant number of independent sources, each containing polylogarithmic min-entropy.

Theorem 1. *There exist constants $c, C > 0$ and a small constant $\beta > 0$ such that for all $n \in \mathbb{N}$ and $k \geq \log^c n$, there exists a polynomial time computable extractor for (n, k, C)-sumset sources, with error $n^{-\Omega(1)}$ and output length k^β.*

1.2 Relations and Applications to Other Sources

Independent Sources

The class of independent sources is clearly a special case of sumset sources. That is, if we view the joint distribution of several independent sources as one source $\mathbf{X}$, then $\mathbf{X}$ is also a sumset

source. Thus, our construction in Theorem 1 also gives an extractor for a constant number of independent sources with polylogarithmic min-entropy. If we can improve the construction and obtain an explicit extractor for $(n, k, 2)$-sumset sources with $k \geq \log^c n$, then this will also match the two source extractors in [CZ15a, Li15b].

Small-Space Sources

As mentioned before, Trevisan and Vadhan [TV00] introduced the problem of constructing seedless extractors for the class of samplable sources, where the weak random source is generated by a computationally bounded algorithm. They constructed explicit extractors for such sources based on strong but plausible complexity-theoretic assumptions. Subsequently, Kamp et al. [KRVZ11] studied the problem of constructing seedless extractors for small-space sources, where the weak source is generated by a small width branching program. We define this model more formally below.

Definition 1.6. *[KRVZ11] A space s source $\mathbf{X}$ is generated by taking a random walk on a branching program of length n and width 2^s, where each edge of the branching program is labelled with a transition probability and a bit. Thus a bit of the source is generated for each step taken on the branching program, and the source X is the concatenation of all the bits.*

As observed in [KRVZ11], the model of small space sources generalizes many previously studied sources, including von Neumann's source of independent coin flips with unknown bias [vN51], the finite Markov chain model studied by Blum [Blu86], a generalization of bit-fixing sources known as symbol-fixing sources [KZ07], and sources consisting of many independent sources. However, the class of affine sources appears not to be related to small space sources.

Using the probabilistic method, one can show that error ε extractors exist for space s sources with min-entropy $k \geq 2s + \log s + O(\log(n/\epsilon))$. However, previously the best known explicit extractor for space s sources is from the work of Kamp et al. [KRVZ11], which requires min-entropy $k \geq \gamma n$ and space $s \leq \gamma^3 n$, where $\gamma > n^{-\delta}$ for some small universal constant δ. In other words, their extractor requires almost linear min-entropy even for sources with space as small as 1, while we know from the probabilistic method that for space $O(\log n)$ sources one can hope to construct extractors for min-entropy $O(\log n)$. In addition, the techniques used in [KRVZ11] start out by reducing to the so called *total-entropy independent sources*, and it can be shown that this reduction has a fundamental bottleneck and cannot possibly go below min-entropy $\sqrt{n}$.

We show how to extract from space s sources when $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, for any constant $\alpha > 0$. Thus for $s = n^{o(1)}$, we only need min-entropy $n^{o(1)}$. This significantly improves previous results in terms of min-entropy requirement, and in particular break the $\sqrt{n}$ min-entropy barrier.

Theorem 2. *For any constant $\alpha > 0$ and for all $n, k, s \in \mathbb{N}$ with $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, there exists a polynomial time computable extractor for space s sources on n bits with min-entropy at least k, with error $n^{-\Omega(1)}$ and output length k^{α}.*

We obtain our result by showing a reduction from the task of extracting from small-space sources to the problem of extracting from sumset sources. We briefly describe the reduction below and refer the reader to Section 3 for more details. Our extractor follows immediately from the reduction.

Note that as observed in [KRVZ11], if we partition a small space source into several blocks, and condition on the event that the branching program generating the source reaches some specific vertices at the end of each block, then the small space source becomes a convex combination

of independent sources. This conditioning reduces the min-entropy of the source, but since the branching program has small width we would expect that there is still much entropy left. However, the problem is that the entropy could now be distributed in these blocks in some arbitrary way, with the only guarantee being a lower bound on the total amount of entropy. This is referred to as a *total entropy source* as in [KRVZ11]. The problem with the approach in [KRVZ11] is that one has to use a fixed partition of the source, so that the blocks can be used as inputs to an extractor for independent sources. This introduced a bottleneck of entropy $\sqrt{n}$, since if the block size is smaller than $\sqrt{n}$ then it could be the case that each block has entropy 1, while if the block size is larger than $\sqrt{n}$ then it could be the case that all entropy is concentrated in just one block.

We get around this obstacle by not relying on a fixed partition of the source. Instead, we show that when the min-entropy satisfies $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, the small space source is actually $2^{-k^{\Omega(1)}}$-close to a convex combination of (n, k^α, C)-sumset sources. On a high level, we show this reduction as follows. We first partition the small space source into some $\ell \gg C$ blocks with $\ell s \ll k$, and we condition on the fixing of the states of the random walk at the end of each block. This leaves us ℓ independent blocks such that their total min-entropy is roughly $k - \ell s$. Now if for some particular fixing, there are at least C blocks with min-entropy at least k^α, then under this fixing the source is an (n, k^α, C)-sumset source. If not, then our key observation is that most of the entropy (indeed, $k - \ell s - \ell k^\alpha = k - o(k)$ entropy) will be concentrated in at most $C - 1$ blocks. Therefore at least one block has min-entropy $(k - o(k))/(C - 1)$. Thus, for this block the entropy rate will be increased by a factor of roughly ℓ/C. We can then fix all other blocks and repeat the argument for this block. Specifically, we further divide the block into ℓ blocks and condition on the fixing of the intermediate states. Then for any particular fixing, either it is an (n, k^α, C)-sumset source or the entropy rate of one block gets increased again by a factor of ℓ/C. We note that the entropy rate cannot be larger than 1, so we know at some point it has to be an (n, k^α, C)-sumset source. Therefore the original source is a convex combination of sumset sources. Notice here the partitions are not fixed, but rather can be different for different fixings of the states.

Interleaved Sources

Raz and Yehudayoff [RY11] introduced a natural generalization of the class of independent sources, which was called *interleaved sources* in a subsequent work by Chattopadhyay and Zuckerman [CZ15b]. We now formally define this class of sources. Let $\sigma : [n] \to [n]$ be any permutation. For any string $w \in \{0,1\}^n$, define the string $s = w_\sigma \in \{0,1\}^n$ such that $s_{\sigma(i)} = w_i$ for $i = 1, \ldots, n$.

Definition 1.7 (Interleaved Sources)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_C$ be independent (n, k)-sources on $\{0,1\}^n$ and let $\sigma : [Cn] \to [Cn]$ be any permutation. Then $\mathbf{Z} = (\mathbf{X}_1 \circ \ldots \circ \mathbf{X}_C)_\sigma$ is an (n, k, C)-interleaved source.*

Besides being a natural generalization of independent sources, the original motivation for studying these sources came from an application found by Raz and Yehudayoff [RY11] in proving lower bounds for arithmetic circuits. Further such extractors give examples of explicit functions with high best-partition communication complexity. Chattopadhyay and Zuckerman [CZ15b] also showed an application to extracting from a generalization of small-space sources where the underlying branching program is any-order.

Using the probabilistic method, one can show that extractors exist for (n, k, C)-interleaved sources with $C = 2$ and $k = O(\log n)$. However the known constructions are far from this in terms of entropy requirement. The construction in [RY11] works for (n, k, C)-interleaved sources for $k > (1 - \delta)n$ and $C = 2$. This was subsequently improved in [CZ15b], where they required one

source with min-entropy $(1-\delta)n$ and the other source with min-entropy $O(\log n)$.

Note that an (n,k,C)-interleaved source is also a special case of an (n,k,C)-sumset source, by naturally extending each source in the definition to have bits 0 in all other positions. Using our extractor for sumset sources, we thus substantially improve previous results in terms of min-entropy requirement. In particular, we obtain explicit extractors that work for the interleaving of a constant number of independent sources, each with polylogarithmic min-entropy.

Theorem 3. *There exist constants $c, C > 0$ and a small constant $\beta > 0$ such that for all n, $k \in \mathbb{N}$ with $k \geq \log^c n$, there exists a polynomial time computable extractor for (n,k,C)-interleaved sources, with error $n^{-\Omega(1)}$ and output length k^β.*

Proof. Suppose $\mathbf{X}$ on Cn is an interleaving of the independent sources $\mathbf{X}_1, \ldots, \mathbf{X}_C$ (each on n bits). Define independent sources $\mathbf{Y}_1, \ldots, \mathbf{Y}_C$, each on Cn bits, such that $\mathbf{Y}_i$ matches $\mathbf{X}$ on the co-ordinates belonging to the source $\mathbf{X}_i$, and $\mathbf{Y}_i$ is fixed to 0 everywhere else. Hence $\mathbf{X} = \sum_1^C \mathbf{Y}_i$ and thus, $\mathbf{X}$ is a (Cn,k,C)-sumset source. The result now follows from Theorem 1. □

We note that, if we can improve our construction and obtain an explicit extractor for $(n,k,2)$-sumset sources with $k \geq \log^c n$, then this will also give an explicit extractor for $(n,k,2)$-interleaved sources with $k \geq \log^c n$.

Affine Sources

An affine source X on n bits with entropy k is the uniform distribution over some unknown affine subspace of dimension k in $\{0,1\}^n$ (viewing $\{0,1\}^n$ as $\mathbb{F}_2^n$[1]). This model generalizes oblivious bit-fixing sources (where some of the bits are uniform and independent, while others are fixed) and thus has received attention for its applications to cryptography. Affine extractors have also been used by Viola [Vio14] to construct extractors for sources generated by NC^0 and AC^0 circuits. Further, good affine extractors imply the best known circuit lower bounds [DK11, FGHK15].

Using the probabilistic method, one can show that affine extractors exist for entropy $k = O(\log n)$. However until recently, the best known explicit constructions for affine extractor was due to Bourgain [Bou07], who using sophisticated techniques from additive combinatorics and gave an explicit extractor for min-entropy at least δn, for any constant δ. This construction was subsequently slightly improved to entropy $n/\sqrt{\log\log n}$ by Yehudayoff [Yeh11] and Li [Li11]. In a very recent work, Li [Li15a] constructed the first explicit affine extractors for polylogarithmic entropy.

We note that an affine source is also a special case of sumset source, since an affine subspace of dimension k can be written as the sum of C affine subspaces of dimension k/C. Thus, as a direct corollary of our extractor for sumset sources, we also obtain extractors for affine sources with polylogarithmic min-entropy, matching the recent work of Li [Li15a].[2]

Corollary 1.8. *There exists a constant $c > 0$ and a small constant $\beta > 0$ such that for all $n, k \in \mathbb{N}$ with $k \geq \log^c n$, there exists a polynomial time computable extractor for affine sources in $\{0,1\}^n$ with entropy k. The extractor has error $n^{-\Omega(1)}$ and output length k^β.*

Proof. Let $\mathbf{X}$ be an affine source with min-entropy k. Let $v_1, \ldots, v_k$ be a basis of $\mathbf{X}$ and b be the shift vector. Let C be the constant in Theorem 1. For $i \in [C]$, define the source $\mathbf{X}_i$ to be the

[1] In general, affine sources can be defined on any field $\mathbb{F}_q$, but in this paper we focus on $\mathbb{F}_2$.

[2] The extractor construction is essentially the same as in [Li15a], but the analysis is different.

uniform distribution on the linear subspace spanned by $v_{(i-1)k/C+1} \ldots, v_{ik/C}$ for $i = 2, \cdots, C$, and define $\mathbf{X}_1$ to be the uniform distribution on the affine subspace spanned by $v_1 \ldots, v_{k/C}$ with shift vector b. Thus $\mathbf{X} = \sum_{j=1}^{C} \mathbf{X}_i$, where each $\mathbf{X}_i$ has min-entropy k/C and the $\mathbf{X}_i$'s are independent. Thus $\mathbf{X}$ is a $(n, k/C, C)$-sumset source, and we can now apply Theorem 1. □

Total Entropy Independent Sources and Somewhere Entropy Independent Sources

As an intermediate model to extract from small space sources, [KRVZ11] introduced the above mentioned *total entropy independent sources*. This is a collection of r independent sources of length ℓ such that the total min-entropy of all r sources is at least k. By the probabilistic method, one can show that error ϵ extractors exist for total min-entropy k independent sources as long as $k \geq \max\{\ell, \log\log(r/\epsilon)\} + \log r + 2\log(1/\epsilon) + O(1)$.[3] Essentially, k can be as small as $\ell + \log r + O(1)$. However, the best known extractors in [KRVZ11] are far from this. Specifically, the extractors there need to have either $k \geq \Omega(r\ell)$ or $k \geq (2^\ell \log r)^C$ for some constant $C > 1$.

We substantially improve these results by constructing a new extractor that only requires min-entropy $O(\ell) + \text{polylog}(r\ell)$, which comes close to the probabilistic bound. In particular, we have

Theorem 1.9. *There exist constants $c, C > 0$ and a small constant $\beta > 0$ such that for all $r, \ell, k \in \mathbb{N}$ with $k \geq C(\ell + \log^c(r\ell))$, there exists a polynomial time computable extractor for r independent sources over $\{0,1\}^\ell$ with total min-entropy k, with error $(r\ell)^{-\Omega(1)}$ and output length k^β.*

To prove the theorem we show the following lemma.

Lemma 1.10. *For any $t, C \in \mathbb{N}$, let $\mathbf{X}_1, \cdots, \mathbf{X}_r \in (\{0,1\}^\ell)^r$ be r independent sources over $\{0,1\}^\ell$ with total min-entropy $k \geq C(\ell + t)$. Then there exists a partition of the r sources into C disjoint subsets $\mathbf{Y}_1, \cdots, \mathbf{Y}_C$ such that each Y_i has min-entropy at least t.*

Proof. We prove the lemma by induction on C. For the case where $C = 1$, one can view the whole set $\mathbf{X}_1, \cdots, \mathbf{X}_r$ as a partition Y_1, and it is clear that Y_1 has min-entropy $k \geq \ell + t > t$. Now suppose the lemma holds for some $C \in \mathbb{N}$, we show that it holds for $C + 1$.

First notice that for two independent sources $\mathbf{X}, \mathbf{Y}$, we have that $H_\infty(\mathbf{X} \circ \mathbf{Y}) = H_\infty(\mathbf{X}) + H_\infty(\mathbf{Y})$. Now, consider the smallest i such that $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_i$ has min-entropy at least t. We know such an i exists because $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_r$ has min-entropy at least $k \geq (C+1)(\ell + t) > t$. Since i is the smallest, we know that $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_{i-1}$ has min-entropy at most t. Note that $\mathbf{X}_i$ has min-entropy at most ℓ, thus $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_i$ has min-entropy at most $t + \ell$. Next, since $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_r$ has min-entropy at least $k \geq (C+1)(\ell+t)$, we know that $\mathbf{X}_{i+1} \circ \cdots \circ \mathbf{X}_r$ has min-entropy at least $k - (t + \ell) = C(t + \ell)$. Now we can apply the induction hypothesis and we see that there exists a partition of $\mathbf{X}_{i+1} \cdots \mathbf{X}_r$ into C disjoint subsets such that each subset has min-entropy at least t. Put in $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_i$ we get $C + 1$ disjoint subsets. □

By setting $t = \log^c(r\ell)$ and combining the lemma with Theorem 1, we immediately obtain Theorem 1.9.

In order to extract from total entropy independent sources, [KRVZ11] actually argues that since the total entropy is at least k, some of the independent sources will have entropy at least k' (the relation between k and k' depends on the number of sources). Therefore, total entropy sources reduce to independent sources where some of them have a certain amount of min-entropy. We call such sources *somewhere entropy independent sources*.

[3] Note that $k > \ell$ is necessary, otherwise the entropy could be contained in just one source, making extraction impossible.

Definition 1.11. *An (n, k, ℓ)-somewhere-C source consists of ℓ independent sources $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$, each on n bits, such that at least C of the $\mathbf{X}_i$'s have min-entropy at least k.*

Note that C here needs to be at least 2. In this context, our extractor for sumset sources from Theorem 1 actually gives an extractor for an (n, k, ℓ)-somewhere-C source with $k \geq \log^c n$ for some constants $C, c > 1$, and outputs $k^{\Omega(1)}$ bits. Note that the number of sources ℓ here is irrelevant since we can just take the sum of the sources and fix any other source that does not have min-entropy k.

In fact, we can use a simpler method to get a slightly stronger result. We show that we can extract from (n, k, ℓ)-somewhere 2 sources for $k = \text{polylog}(n)$ and any integer ℓ (with the extractor running in time $\text{poly}(n, \ell)$).

Theorem 1.12. *There exists a constant $c > 0$ and a small constant $\beta > 0$ such that for all $n, k, \ell \in \mathbb{N}$ with $k \geq \log^c n$, there exists an extractor computable in time $\text{poly}(n, \ell)$ for (n, k, ℓ)-somewhere-2 sources, with error $n^{-\Omega(1)}$ and output length $\Omega(k)$.*

We provide the proof of Theorem 1.12 in Section 7.

1.3 Outline of Constructions

We now give an informal description of our extractor for sumset sources. On a very high level, our extractor follows the same spirit of recent works on extractors [CZ15a, Li15b, Li15a]. That is, we first convert our sumset source into a special non-oblivious bit-fixing source, which is a distribution on $N = n^{O(1)}$ bits such that $N - N^\delta$ bits are k^α-wise independent, for some constants $0 < \delta, \alpha < 1$. We will then apply extractors for this source constructed in [CZ15a, Li15b].

To obtain such a non-oblivious bit-fixing source, it suffices to use two *independent* sources as shown in [Li15c, CZ15a]. More specifically, if we have a somewhere random source[4] with N rows such that $N - N^\delta$ rows are uniform, then by taking another independent source and using techniques based on alternating extraction [DP07, DW09, Li13a, Li15c, Coh15, CGL15], one can obtain the desired non-oblivious bit-fixing source.

However, in our case we do not have independent sources, but rather the sum of independent sources. Our first observation is that, since sum is a linear operation, in this case alternating extraction is still possible, as long as in each step the seeded extractor used is a *linear seeded extractor*. Informally, a linear seeded extractor is a strong seeded extractor with the extra property such that for any fixing of the seed, the output of the extractor is a linear function of the source. This property has been used several times in the literature of extractors for affine sources, and in particular alternating extraction between linearly correlated sources has been used by Li in his recent construction of affine extractors [Li15a].

Now the problem is how to obtain the somewhere random source. The standard way is to use a seeded extractor with seed length $O(\log n)$ (so that to keep the running time polynomial in n) and try all possible values of the seed. Each seed will give an output and we can then concatenate the output to form a matrix. This does indeed give us a somewhere random source, however there are now two problems. First, we cannot just use any seeded extractor with seed length $O(\log n)$. This is because we need to apply the seeded extractor to the sum of several independent sources, and we need to keep the "sum" structure carefully for the purpose of alternating extraction later. If we just use any seeded extractor, then after applying the extractor the "sum" structure may not be preserved. Therefore, here again we need to use a linear seeded extractor. Luckily, we do have linear seeded extractors with seed length $O(\log n)$, due to a construction in [Li15a].

[4] A somewhere random source is a matrix of random variables such that at least one row is uniform.

Second, just doing this is not enough, since the error of the somewhere random source is not good enough for our purpose. Specifically, in order to apply the extractor for non-oblivious bit-fixing source we need the error to be negligibly small, while the error we obtained from a seeded extractor with seed length $O(\log n)$ is only polynomially small. Note this is different from the affine extractor construction in [Li15a], as in the case of affine sources one can show that if we use a linear seeded extractor, then most of the rows in the somewhere random source actually have error 0. However for general weak random sources the best error one can hope for (even with a linear seeded extractor) is $1/\text{poly}(n)$ if the seed length is $O(\log n)$.

To get around this, both [Li15c] and [CZ15a] used a sampling method. Specifically, they first used an extractor (or a non-malleable extractor) with large seed length to achieve small error from one source, and then use another independent source to sample from the rows of the somewhere random source to bring down the number of rows. The first idea would be to try the same idea here in our construction. That is, if $\mathbf{X}$ is the sum of two independent sources, then one can take two linear seeded extractors $\text{Ext}_1, \text{Ext}_2$ such that Ext_1 has large seed length, Ext_2 has seed length $O(\log n)$ and output length the same as the seed length of Ext_1, and compute $\text{Ext}_1(\mathbf{X}, \text{Ext}_2(\mathbf{X}, r))$ for every possible choice r of Ext_2's seed. However, the problem now is that the sampling procedure becomes correlated, and even with linear seeded extractors we do not know how to analyze it.

We thus turn to another approach, used by Li in his multi-source extractor [Li13a]. The idea is that, assume that $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_C$ is the sum of some constant C number of independent sources (instead of just two independent sources). Then if we apply a linear seeded extractor to $\mathbf{X}$, by the property of the extractor for every fixed seed the output will also be the XOR of C independent outputs from each $\mathbf{X}_i$. If every output is ϵ-close to uniform for some error ϵ, then the error after the XOR will be reduced to roughly ϵ^C. Thus, if we take C to be a large enough constant, this error will be much smaller than $1/N$ where N is the number of rows in the somewhere random source. At this point we can use a union-bound type argument to show that the somewhere random source is actually $N\epsilon^C = 1/\text{poly}(n)$-close to another somewhere random source where a large fraction of the rows are *truly uniform*. Thus we can switch to the new somewhere random source and only introduce an error of $1/\text{poly}(n)$.

Now, if we have one more independent source, that is $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_C + \mathbf{X}_{C+1}$, then we can apply alternating extraction between the somewhere random source and $\mathbf{X}$ itself, using linear seeded extractors. An analysis similar to that of [Li15c, Coh15, Li15a] will show that at the end we will obtain a k^α-wise independent non-oblivious bit-fixing source as desired.

1.4 Open Problems

Our paper leaves open many natural questions. The most obvious question is to improve the output length and error of our extractors. A further question is, can we improve the construction to give an extractor for the sum of two independent sources with polylogarithmic min-entropy? This will imply two-source extractors and extractors for the interleaving of two independent sources, as well as improving some of the parameters in this paper. Again, a natural approach is to adjust the sampling approach in [Li15c, CZ15a] to the case of sumset sources. Also, our extractor for small space sources is not optimal, and there is still much room for improvement. Can we find a better reduction from small space sources to sumset sources?

As shown in our paper, the model of sumset source and the extractors for this kind of source seem to be very powerful, in the sense that it generalizes many previously studied sources and gives improved extractors. Can we use our construction to give improved extractors for other sources? Finally, a more general model would be to look at other functions of independent sources, rather

than just looking at the sum. Indeed, this is a strict generalization of samplable sources, since any source generated by some class of function where the input is the uniform string can be viewed as the function applied to several independent sources.

1.5 Organization

In Section 3, we provide a reduction from small-space sources to sumset sources. We use Section 2 to introduce some necessary preliminaries and prior work. In Section 4, we prove one of our main technical ingredients for designing extractors for sumset sources. In particular, we extend the method of alternating extraction to correlated sources. Next, in Section 5, we show a method of breaking correlations among random variables using another correlated source. As discussed above, the work of Cohen [Coh15] showed how to do this but required an additional independent source. Using these components and other explicit extractors from prior work, we provide explicit extractors for sumset sources in Section 6. Finally, we prove Theorem 1.12 in Section 7.

2 Preliminaries

We use $\mathbf{U}_m$ to denote the uniform distribution on $\{0,1\}^m$.
For any integer $t > 0$, $[t]$ denotes the set $\{1, \ldots, t\}$.
For a string y of length n, and any subset $S \subseteq [n]$, we use y_S to denote the projection of y to the coordinates indexed by S.
We use bold capital letters for random variables and samples as the corresponding small letter, e.g., $\mathbf{X}$ is a random variable, with x being a sample of $\mathbf{X}$.

2.1 Linear Seeded Extractors

Definition 2.1. *A function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* (k, ϵ)*-seeded extractor if for any source* $\mathbf{X}$ *of min-entropy* k, $|\mathrm{Ext}(\mathbf{X}, \mathbf{U}_d) - \mathbf{U}_m| \leq \epsilon$. Ext *is called a strong seeded extractor if* $|(\mathrm{Ext}(\mathbf{X}, \mathbf{U}_d), \mathbf{U}_d) - (\mathbf{U}_m, \mathbf{U}_d)| \leq \epsilon$, *where* $\mathbf{U}_m$ *and* $\mathbf{U}_d$ *are independent. Further, if for each* $s \in \mathbf{U}_d$, $\mathrm{Ext}(\cdot, s) : \{0,1\}^n \to \{0,1\}^m$ *is a linear function, then* Ext *is called a linear seeded extractor.*

We use explicit constructions of linear seeded extractors.

Theorem 2.2 ([Tre01] [RRV02])**.** *For every* $n, k, m \in \mathbb{N}$ *and* $\epsilon > 0$, *with* $m \leq k \leq n$, *there exists an explicit strong linear seeded extractor* $\mathrm{LExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for min-entropy* k *and error* ϵ, *where* $d = O\left(\frac{\log^2(n/\epsilon)}{\log(k/m)}\right)$.

A drawback of the above construction is that the seeded length is $\omega(\log n)$ for sub-linear min-entropy. A recent construction of Li [Li15a] achieves $O(\log n)$ seed length for even polylogarithmic min-entropy.

Theorem 2.3 ([Li15a])**.** *There exists a constant* $c > 1$ *such that for every* $n, k \in \mathbb{N}$ *with* $c \log^8 n \leq k \leq n$ *and any* $\epsilon \geq 1/n^2$, *there exists a polynomial time computable linear seeded extractor* $\mathrm{LExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for min-entropy* k *and error* ϵ, *where* $d = O(\log n)$ *and* $m \leq \sqrt{k}$.

2.2 2-Source Extractors

We use explicit constructions of 2-source extractors from the recent work of Chattopadhyay and Zuckerman [CZ15a], with subsequent improvement of the output length by Li [Li15b].

Theorem 2.4 ([CZ15a, Li15b]). *There exists a constant $\lambda > 0$ such that for all $n \in \mathbb{N}$, there exists a polynomial time computable 2-source extractor $2\text{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$, $m = \Omega(k)$, for min-entropy $k > \log^\lambda(n)$ and error $n^{-\Omega(1)}$.*

2.3 Extractors for Non-Oblivious Bit-fixing Sources

Definition 2.5. *A distribution $\mathcal{D}$ on n bits is t-wise independent if the restriction of $\mathcal{D}$ to any t bits is uniform. Further $\mathcal{D}$ is a (t,ϵ)-wise independent distribution if the distribution obtained by restricting $\mathcal{D}$ to any t coordinates is ϵ-close to uniform.*

Definition 2.6. *A source $\mathbf{X}$ on $\{0,1\}^n$ is called a (q,t)-non-oblivious bit-fixing source if there exists a subset of coordinates $Q \subseteq [n]$ of size at most q such that the joint distribution of the bits indexed by $\overline{Q} = [n] \setminus Q$ is t-wise independent. The bits in the coordinates indexed by Q are allowed to arbitrarily depend on the bits in the coordinates indexed by $\overline{Q}$.*

If the joint distribution of the bits indexed by $\overline{Q}$ is (t,γ)-wise independent then $\mathbf{X}$ is said to be a (q,t,γ)-non-oblivious bit-fixing source.

We use extractors for (q,t,γ)-non-oblivious bit fixing sources constructed in [CZ15a] with subsequent improvement in the output length in [Li15b].

Theorem 2.7 ([CZ15a, Li15b]). *There exists a constant λ and a small constant $\alpha > 0$ such that for any constant $\delta > 0$, and for all $n, q, t, m \in \mathbb{N}$ satisfying $q \leq n^{1-\delta}, t \geq \log^\lambda(n), m = t^\alpha$, there exists an explicit extractor $\text{bitExt} : \{0,1\}^n \to \{0,1\}^m$ for the class of (q,t,γ)-non-oblivious bit-fixing sources with error $n^{-\Omega(1)}$, where $\gamma \leq 1/n^{t+1}$.*

2.4 Conditional Min-Entropy

Definition 2.8. *The average conditional min-entropy of a source $\mathbf{X}$ given a random variable $\mathbf{W}$ is defined as*

$$\widetilde{H}_\infty(\mathbf{X}|\mathbf{W}) = -\log\left(\mathbf{E}_{w\sim W}\left[\max_x \Pr[\mathbf{X} = x|\mathbf{W} = w]\right]\right) = -\log\left(\mathbf{E}\left[2^{-H_\infty(\mathbf{X}|\mathbf{W}=w)}\right]\right).$$

We recall some results on conditional min-entropy from the work of Dodis et al. [DORS08].

Lemma 2.9 ([DORS08]). *For any $\epsilon > 0$, $\mathbf{Pr}_{w\sim\mathbf{W}}\left[H_\infty(\mathbf{X}|\mathbf{W} = w) \geq \widetilde{H}_\infty(\mathbf{X}|\mathbf{W}) - \log(1/\epsilon)\right] \geq 1 - \epsilon$.*

Lemma 2.10 ([DORS08]). *If a random variable $\mathbf{Y}$ has support of size 2^ℓ, then $\widetilde{H}_\infty(\mathbf{X}|\mathbf{Y}) \geq H_\infty(\mathbf{X}) - \ell$.*

We require extractors that can extract uniform bits when the source only has sufficient conditional min-entropy.

Definition 2.11. *A (k,ϵ)-seeded average case seeded extractor $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ for min-entropy k and error ϵ satisfies the following property: For any source $\mathbf{X}$ and any arbitrary random variable $\mathbf{Z}$ with $\tilde{H}_\infty(\mathbf{X}|\mathbf{Z}) \geq k$,*

$$\text{Ext}(\mathbf{X}, \mathbf{U}_d), \mathbf{Z} \approx_\epsilon \mathbf{U}_m, \mathbf{Z}.$$

It was shown in [DORS08] that any seeded extractor is also an average case extractor.

Lemma 2.12 ([DORS08]). *For any $\delta > 0$, if Ext is a (k,ϵ)-seeded extractor, then it is also a $(k + \log(1/\delta), \epsilon + \delta)$-seeded average case extractor.*

2.5 Some Probability Lemmas

The following result on min-entropy was proved by Maurer and Wolf [MW97].

Lemma 2.13. *Let $\mathbf{X}, \mathbf{Y}$ be random variables such that the random variable $\mathbf{Y}$ takes at ℓ values. Then*

$$\mathbf{Pr}_{y \sim \mathbf{Y}}\left[H_\infty(\mathbf{X}|\mathbf{Y}=y) \geq H_\infty(\mathbf{X}) - \log \ell - \log\left(\frac{1}{\epsilon}\right)\right] > 1-\epsilon.$$

Lemma 2.14 ([BIW06])**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$ be independent random variables on $\{0,1\}^m$ such that $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$. Then, $|\sum_{i=1}^{\ell} \mathbf{X}_i - \mathbf{U}_m| \leq \epsilon^\ell$.*

3 A Reduction from Small-Space Sources to Sumset Sources

In this section, we show that a small-space source is close to a convex combination of sumset sources. The idea is argue that either partitioning the source leads to a sumset source or results in increase in min-entropy rate of one of the partitions. Thus by repeating this argument, it must be that at some point we reach a sumset source, since otherwise we end up with a source with min-entropy rate more than 1.

Lemma 3.1. *For any constant $\alpha > 0$ and any constant integer $C \geq 2$, any space s source on n bits with min-entropy $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$ is $2^{-k^{\Omega(1)}}$-close to a convex combination of (n, k', C)-sumset sources, where $k' = k^\alpha$.*

Proof. Let $\ell = k^{\alpha/2}, \epsilon_1 = 2^{-k^\alpha}, k_{th} = k^\alpha$ be fixed parameters that we set with foresight. Let $\mathbf{X}$ be a space s source on n bits with min-entropy at least k. We partition $\mathbf{X}$ into ℓ equi-sized blocks of length $n_1 = n/\ell$. Let $\mathbf{X}_i$, denote the i'th block where $i \in [\ell]$ (thus $\mathbf{X}_i$ is a source on n/ℓ bits). We now condition on the initial state of small-space branching program at each of these blocks, and let k_i denote the min-entropy in $\mathbf{X}_i$ after this conditioning. Observe that $\mathbf{X}_i$'s are now independent sources. It follows from Lemma 2.13 that with probability at least $1-\epsilon_1$,

$$\sum_{i=1}^{\ell} k_i \geq k - \ell s - \log(1/\epsilon_1). \tag{1}$$

Consider any such good fixing of the states such that the above inequality holds. The proof now goes via analysing two cases. Since we iterate this argument, each time with a new source, let $\mathbf{X}^1 = \mathbf{X}$ and $k_{(1)} = k$.

Case 1: $|\{i \in [\ell] : k_i \geq k_{th}\}| \geq C$. The proof is direct in this case. For simplicity, suppose $\mathbf{X}_1, \ldots, \mathbf{X}_C$ each have min-entropy at least k^α. We fix the sources $\mathbf{X}_{C+1}, \ldots, \mathbf{X}_\ell$. Now, for each $i \in [C]$, define the source $\mathbf{Y}_i$ on n bits whose projection onto the the i'th block is $\mathbf{X}_i$ and the rest of the co-ordinates are fixed to 0. It follows that $\mathbf{X} = \eta + \sum_{i=1}^{C} \mathbf{Y}_i$ (for some constant string $\eta \in \{0,1\}^n$) and hence is a (n, k', C)-sumset source. Thus $\mathbf{X}$ is at distance at most ϵ_1 from a convex combination of such sumset sources.

Case 2: $|\{i \in [\ell] : k_i \geq k_{th}\}| < C$. Using (1), it follows that there exists distinct $C-1$ partitions, say $i_1, \ldots, i_{C-1}$ such that

$$\sum_{j=1}^{C-1} k_{i_j} \geq k_{(1)} - \ell(s + k^\alpha) - \log(1/\epsilon_1).$$

Thus, by an averaging argument, it follows that there exists some $j \in [C-1]$, such that

$$k_{i_j} \geq \frac{k_{(1)} - \ell(s+k^\alpha) - \log(1/\epsilon_1)}{C-1}.$$

Hence the source $\mathbf{X}_{i_j}$ (on $n_1 = n/\ell$ bits) has min-entropy rate

$$\frac{k_{(1)} - \ell(s+k^\alpha) - \log(1/\epsilon_1)}{C-1} \cdot \frac{\ell}{n}$$

Thus, using the fact that $k_{(1)} > (sk^{\alpha/2} + 2k^{\alpha+\alpha})^{1+\alpha}$, the min-entropy rate of $\mathbf{X}_{i_j}$ is at least $\frac{k_{(1)}\ell}{2nC}$, and hence

$$\frac{H_\infty(\mathbf{X}_{i_j})}{n_1} \geq \frac{\ell}{2C} \cdot \frac{H_\infty(\mathbf{X}^1)}{n}.$$

We now repeat the argument (i.e, analyzing the Cases 1 and 2) with $\mathbf{X}^1$ replaced by $\mathbf{X}^2 = \mathbf{X}_{i_j}$ (and we fix all other sources). However, for different iterations of the argument, we do not change values of the parameters ℓ, ϵ, k_{th}, and they are fixed to k^α, 2^{-k^α} and k^α respectively, where $k = H_\infty(\mathbf{X})$.

Suppose, if possible, that for h iterations of this argument, each time we end up in Case 2. Thus, we now have a source $\mathbf{X}^h$ on n/ℓ^h bits with min-entropy rate at least $(\frac{\ell}{2C})^h \cdot \frac{k}{n}$. To derive a contradiction using the fact that the min-entropy rate is at most 1, we require

- $\left(\frac{\ell}{2C}\right)^h \cdot \frac{k}{n} > 1$,
- $\frac{n}{\ell^h} \geq k^\alpha$
- $\frac{k}{(2C)^h} > (sk^{\alpha/2} + 2k^{2\alpha})^{1+\alpha}$.

(The first condition is to ensure that the min-entropy rate is more than 1, the second condition ensures that the length of the source $\mathbf{X}^h$ is large enough and finally the third condition is a lower bound the min-entropy of $\mathbf{X}^h$, which is required when we apply our argument on $\mathbf{X}^{h-1}$.)

Pick $h = 1 + \frac{\log n - \log k}{\log \ell - \log(2C)}$. It is easy to check that the first condition holds. Further the second and third conditions follow from the fact that $k > s^{1+10\alpha} 2^{\log^{0.5+\alpha}(n)}$. Thus, it must be that in at most h iterations of the argument, we are in Case 1 and hence $\mathbf{X}$ is close to a convex combination of (n, k', C)-sumset sources. We note that the statistical distance to the convex combination is at most $O\left(\epsilon_1 \frac{\log n}{\log k}\right)$. □

4 Alternating Extraction between Correlated Sources

The method of alternating extraction was introduced by Dziembowski and Pietrzak [DP07]. Since its introduction, this technique has found applications in a variety of extractor constructions [DW09, Li13a, Li15c, Coh15, CGL15, Li15a]. In this section we extend this method to the situation when the sources playing the alternating extraction game are correlated.

We recall the method of alternating extraction. Assume that there are two parties, Quentin with a source $\mathbf{Q}$ and a uniform seed $\mathbf{S}_0$, and Wendy with a source $\mathbf{W}$. The protocol is an interactive process between Quentin and Wendy, and starts off with Quentin sending the seed $\mathbf{S}_0$ to Wendy. Wendy uses $\mathbf{S}_0$ and a strong seeded extractor Ext_w to extract a seed $\mathbf{R}_0$ using $\mathbf{W}$, and sends $\mathbf{R}_0$

back to Quentin. This constitutes a round of the alternating extraction protocol. In the next round, Quentin uses a strong extractor Ext_q to extract a seed $\mathbf{S}_1$ from $\mathbf{Q}$ using $\mathbf{S}_0$, and sends it to Wendy and so on. The protocol is run for h steps, where h is an input parameter. Thus, the following sequence of random variables is generated:

$$\mathbf{S}_0, \mathbf{R}_0 = \mathrm{Ext}_w(\mathbf{S}_0), \mathbf{S}_1 = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_0), \ldots, \mathbf{S}_u = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_{h-1}), \mathbf{R}_h = \mathrm{Ext}_w(\mathbf{W}, \mathbf{S}_h).$$

Look-Ahead Extractor: We define the following look-ahead extractor:

$$\mathrm{laExt}(\mathbf{W}, (\mathbf{Q}, \mathbf{S}_0)) = \mathbf{R}_1, \ldots, \mathbf{R}_h.$$

We now prove a lemma establishing a property of the alternating extraction protocol similar in spirit to many of the previous works that used this technique. However, in all previous uses of alternating extraction (apart from [Li15a]), the sources $\mathbf{Q}$ and $\mathbf{W}$ in the protocol are assumed to be independent. We show that such a protocol can be run between sources $\mathbf{W} = \mathbf{X} + \mathbf{Z}$ and $\mathbf{Q} = \mathbf{Y}$, where $\mathbf{Y}$ and $\mathbf{Z}$ are arbitrarily correlated and $\mathbf{X}$ is independent of $(\mathbf{Y}, \mathbf{Z})$. We now formally state this below.

Lemma 4.1. *For any $\epsilon > 0$ and any integers $n_1, n_2, k, k_1, t, d, h$ satisfying $k_1 \geq k + 2(t+1)d(h+1) + \log(1/\epsilon)$ and $n_2 \geq k + 2(t+1)d(h+1) + \log(1/\epsilon)$, let*

- *$\mathbf{X}$ be an (n_1, k_1)-source, $\mathbf{Y} = \mathbf{U}_{n_2}$ and $\mathbf{Z}$ be a random variable on n_1 bits.*
- *$\mathbf{Y}^1, \ldots, \mathbf{Y}^t$ be random variables on n_2 bits each, such that X is independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.*
- *$\mathbf{S}_0 = \mathrm{Slice}(\mathbf{Y}, d)$ and for $i \in [t]$, $\mathbf{S}_0^i = \mathrm{Slice}(\mathbf{Y}^i, d)$.*
- *$\mathrm{LExt}_1 : \{0,1\}^{n_1} \times \{0,1\}^d \to \{0,1\}^d$ and $\mathrm{LExt}_2 : \{0,1\}^{n_2} \times \{0,1\}^d \to \{0,1\}^d$ be (k, ϵ)-strong linear seeded extractors.*
- *$\mathrm{laExt}(\mathbf{X} + \mathbf{Z}, (\mathbf{Y}, \mathbf{S}_0)) = \mathbf{R}_1, \ldots, \mathbf{R}_h$, and for $i \in [t]$, $\mathrm{laExt}(\mathbf{X} + \mathbf{Z}, (\mathbf{Y}^i, \mathbf{S}_0^i)) = \mathbf{R}_1^i, \ldots, \mathbf{R}_h^i$, where laExt is executed with the linear seeded extractors $\mathrm{LExt}_1, \mathrm{LExt}_2$ for h rounds.*
- *$\mathbf{R}_{j,\mathbf{X}} = \mathrm{LExt}_1(\mathbf{X}, \mathbf{S}_j)$ and $\mathbf{R}_{j,\mathbf{Z}} = \mathrm{LExt}_1(\mathbf{Z}, \mathbf{S}_j)$, $j \in [0, h]$.*

Then,

1. *for any $j \geq 0$,*

$$\begin{aligned} &\mathbf{S}_j, \{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}_g^i : g \in [0, j-1], i \in [t]\}, \{\mathbf{R}_g : g \in [0, j-1]\}, \\ &\qquad \{\mathbf{R}_g^i : g \in [0, j-1], i \in [t]\} \\ \approx_{(4j+2)\epsilon}\ &\mathbf{U}_d, \{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}_g^i : g \in [0, j-1], i \in [t]\}, \{\mathbf{R}_g : g \in [0, j-1]\}, \\ &\qquad \{\mathbf{R}_g^i : g \in [0, j-1], i \in [t]\}. \end{aligned}$$

2. *for any $j \geq 0$, conditioned on $\{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}_g^i : g \in [0, j-1], i \in [t]\}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}_g^i : g \in [0, j-1], i \in [t]\}$,*
   - *$\mathbf{X}$ is independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.*
   - *$\mathbf{S}_j$ and $\{\mathbf{S}_j^i : i \in [t]\}$ are deterministic functions of $\mathbf{Y}$.*

- $\mathbf{X}$ *has conditional min-entropy at least* $k + (t+1)d(h+1-j) + \log(1/\epsilon)$ *and* $\mathbf{Y}$ *has conditional min-entropy at least* $k + 2(t+1)d(h+1-j) + \log(1/\epsilon)$.

3. *for any* $j \geq 0$,

$$\begin{aligned} &\mathbf{R}_j, \mathbf{R}_{j,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-1], i \in [t]\}, \\ &\qquad \{\mathbf{S}_g : g \in [0, j]\}, \{\mathbf{S}^i_g : g \in [0, j], i \in [t]\}, \mathbf{Y}, \{\mathbf{Y}^i : i \in [t]\}, \mathbf{Z} \\ \approx_{4(j+1)\epsilon}\ &\mathbf{U}_d, \mathbf{R}_{j,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-1], i \in [t]\}, \\ &\qquad \{\mathbf{S}_g : g \in [0, j]\}, \{\mathbf{S}^i_g : g \in [0, j], i \in [t]\}, \mathbf{Y}, \{Y^i : i \in [t]\}, \mathbf{Z}. \end{aligned}$$

4. *for any* $j \geq 0$, *conditioned on* $\mathbf{R}_{j,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-1], i \in [t]\}, \{\mathbf{S}_g : g \in [0, j]\}, \{\mathbf{S}^i_g : g \in [0, j], i \in [t]\}$,

   - $\mathbf{X}$ *is independent of* $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
   - $\mathbf{R}_j$ *and* $\{\mathbf{R}^i_j : i \in [t]\}$ *are deterministic function of* $\mathbf{X}$.
   - $\mathbf{X}$ *has conditional min-entropy at least* $k + (t+1)d(h+1-j) + \log(1/\epsilon)$ *and* $\mathbf{Y}$ *has conditional min-entropy at least* $k + 2(t+1)d(h-j) + \log(1/\epsilon)$.

Proof. We prove the lemma by induction on j. The validity of the lemma when $j = 0$ is direct. Thus, suppose that the lemma holds for $j-1$ for some $j \in [h]$ and we prove it for j.

Fix the following random variables:

$$\begin{aligned} &\mathbf{R}_{j-1,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-2]\}, \{\mathbf{R}^i_{j-1,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-2], i \in [t]\}, \\ &\qquad \{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}^i_g : g \in [0, j-1], i \in [t]\}. \end{aligned}$$

By induction hypothesis, it follows that

- $\mathbf{R}_{j-1}$ is $4j\epsilon$-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{X}$.
- $\mathbf{Y}$ has conditional min-entropy $k + 2(t+1)d(h+1-j) + \log(1/\epsilon)$ and is independent of $\mathbf{X}$.
- $\mathbf{X}$ has conditional min-entropy $k + (t+1)d(h+2-j) + \log(1/\epsilon)$.

Since $\mathbf{S}_j = \mathrm{LExt}_2(\mathbf{Y}, \mathbf{R}_{j-1})$, it follows by Lemma 2.12 that $\mathbf{S}_j$ is $(4j+2)\epsilon$-close to $\mathbf{U}_d$ on average conditioned on $\mathbf{R}_{j-1}$. Thus we fix $\mathbf{R}_{j-1}$ and observe that $\mathbf{S}_j$ is now a deterministic function of $\mathbf{Y}$. Next we fix $\{\mathbf{R}^i_{j-1} : i \in [t]\}$ observing that, by induction hypothesis, they are deterministic functions of $\mathbf{X}$ and hence does not affect $\mathbf{S}_j$. As a result of this fixing, $\{\mathbf{S}^i_j : i \in [t]\}$ is now a deterministic function of $\mathbf{Y}$, and further $\mathbf{X}$ remains independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. We note that all the random variables fixed in this step are deterministic functions of $\mathbf{X}$. Thus after these fixings, by Lemma 2.10 and induction hypothesis, the conditional entropy of $\mathbf{X}$ is at least $k + (t+1)d(h+2-j) - (t+1)d + \log(1/\epsilon) = k + (t+1)d(h+1-j) + \log(1/\epsilon)$. This concludes the proof of (1) and (2).

We now prove (3) and (4). We continue to condition on the random variables that we have fixed so far in our proof. We have,

- $\mathbf{S}_j$ is $(4j+2)\epsilon$-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{Y}$,
- $\mathbf{X}$ has average conditional min-entropy at least $k + (t+1)(h+1-j) + \log(1/\epsilon)$ and is independent of $\mathbf{Y}$,

- $\mathbf{Y}$ has conditional min-entropy $k + 2(t+1)d(h+1-j) + \log(1/\epsilon)$.

Thus, it follows by Lemma 2.12 that $\mathbf{R}_{j,\mathbf{X}} = \mathrm{LExt}_1(\mathbf{X}, \mathbf{S}_j)$ is $4(j+1)\epsilon$-close to $\mathbf{U}_d$ on average conditioned on $\mathbf{S}_j$. We fix $\mathbf{S}_j$ and note that $\mathbf{R}_{j,\mathbf{X}}$ is now a deterministic function of $\mathbf{X}$. Next, we fix $\mathbf{R}_{j,\mathbf{Z}}$ which is now a deterministic function of $\mathbf{Z}$ and hence does not affect $\mathbf{R}_{j,\mathbf{X}}$. Since LExt_1 is linear seeded, it follows that $\mathbf{R}_j = \mathbf{R}_{j,\mathbf{X}} + \mathbf{R}_{j,\mathbf{Z}}$ and $\mathbf{R}^i_j = \mathbf{R}^i_{j,\mathbf{X}} + \mathbf{R}^i_{j,\mathbf{Z}}$. Thus $\mathbf{R}_j$ is ϵ_j-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{X}$. We now fix $\{\mathbf{S}^i_j : i \in [t]\}$ which is a deterministic function of $\mathbf{Y}$, and next fix $\{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}$ which is a deterministic function of $\mathbf{Z}$. Thus, these additional fixings do not affect $\mathbf{R}_j$. Finally observe that $\mathbf{X}$ remains independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. We note that all the random variables fixed in this step are deterministic functions of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Thus after these fixings, by Lemma 2.10, the conditional entropy of $\mathbf{Y}$ is at least $k + 2(t+1)d(h+1-j) - 2(t+1)d + \log(1/\epsilon) = k + 2(t+1)d(h-j) + \log(1/\epsilon)$. This concludes the proof of induction and hence the lemma follows. □

5 Breaking Correlations Using another Correlated Source

Let $\mathbf{Y}^1, \ldots, \mathbf{Y}^t$ be correlated random variables. Using the method of alternating extraction in clever ways, works by Li [Li13a] and Cohen [Coh15] gave two alternate ways of breaking correlations of these random variables using an additional independent source $\mathbf{X}$.

We show that it is possible to break the correlations even by using an additional correlated source of the form $\mathbf{X} + \mathbf{Z}$, assuming $\mathbf{X}$ is independent of $\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$ (and $\mathbf{Z}$ is allowed to have arbitrary correlations with $\mathbf{Y}^1, \ldots, \mathbf{Y}^t$).

Our idea is to adapt the method from [Coh15] to our setting. One key change is to use linear seeded extractors in the alternating extraction steps. This allows us to use the result from Section 4 on alternating extraction between correlated sources. Our proof technique is similar to most proofs that use alternating extraction and goes via careful conditioning of random variables.[5]

Algorithm 1 uses alternating extraction in a flip-flop way (introduced in [Coh15]). Algorithm 2 chains together several flip-flop steps and is called a local correlation breaker [Coh15]. As discussed above, we instantiate these functions with linear seeded extractors.

Algorithm 1: flip-flop(y^i, y^i_j, w, b)

Input: Bit strings $y^i, y^i_j, w = x + z$ of length n_1, n_2, n_1 respectively, and a bit b.

Output: Bit string y^i_{j+1} of length n_2.

Subroutines: Let $\mathrm{LExt}_1 : \{0,1\}^{n_1} \times \{0,1\}^d \to \{0,1\}^d$, $\mathrm{LExt}_2 : \{0,1\}^{n_2} \times \{0,1\}^d \to \{0,1\}^d$ be (k, ϵ)-strong linear seeded extractors. Let $\mathrm{LExt}_3 : \{0,1\}^{n_1} \times \{0,1\}^d \to \{0,1\}^{n_2}$ be a (k_2, ϵ)-strong linear seeded extractor.

Let $\mathrm{laExt} : \{0,1\}^{n_1} \times \{0,1\}^{n_2+d} \to \{0,1\}^{2d}$ be a look-ahead extractor for an alternating extraction protocol run for 2 rounds using $\mathrm{LExt}_1, \mathrm{LExt}_2$ as the seeded extractors.

1 Let $s^i_{0,j} = \mathrm{Slice}(y^i_j, d)$, $\mathrm{laExt}(w, (y^i_j, s^i_{0,j})) = r^i_{0,j}, r^i_{1,j}$

2 Let $\overline{y^i_{1,j}} = \mathrm{LExt}_3(y^i, r^i_{b,j})$

3 Let $\overline{s^i_{0,j}} = \mathrm{Slice}(\overline{y^i_{1,j}}, d)$, $\mathrm{laExt}(w, (\overline{y^i_{1,j}}, \overline{s^i_{0,j}})) = \overline{r^i_{0,j}}, \overline{r^i_{1,j}}$

4 Output $y^i_{j+1} = \mathrm{LExt}_3(y^i, \overline{r^i_{1-b,j}})$

[5] It is also possible to use the method of [Li13a] but using [Coh15] gives us slightly better parameters.

Algorithm 2: LCB(y^i, w, id)

Input: Bit strings $y^i, w = x + z, id$ of length n_1, n_1, h respectively.
Output: Bit string y_{h+1} of length n_2.

Let $y^i_1 = \text{Slice}(y, n_2)$
for $j = 1$ *to* h **do**
  $y^i_{j+1} = \text{flip-flop}(y^i, y^i_j, w, id[j])$
end
Output y^i_{h+1}.

The following is the main result of this section.

Theorem 5.1. *For any $\epsilon > 0$ and any integers $n_1, n_2, k, k_1, t, d, h$ satisfying $k_1 \geq k + 8tdh + \log(1/\epsilon)$, $n_2 \geq k + 3td + \log(1/\epsilon)$, $n_1 \geq k + 10tdh + (4ht + 1)n_2 + \log(1/\epsilon)$, let*

- $\mathbf{X}$ *be an (n_1, k_1)-source, $\mathbf{Y}^1 = \mathbf{U}_{n_1}$ and $\mathbf{Z}, \mathbf{Y}^2, \ldots, \mathbf{Y}^t$ be random variables on n_1 bits each, such that X is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.*
- *$id^1, \ldots, id^t$ be bit strings of length h such that for each $i \in [t]$, $id^1 \neq id^i$.*
- *$\mathbf{Y}^i_{h+1} = \text{LCB}(\mathbf{Y}, \mathbf{X} + \mathbf{Z}, id^i)$ for $i \in [t]$ where* LCB *is the function computed by Algorithm 2.*

Then,

$$\mathbf{Y}^1_{h+1}, \mathbf{Y}^2_{h+1}, \ldots, \mathbf{Y}^t_{h+1} \approx_{O(h\epsilon)} \mathbf{U}_{n_2}, \mathbf{Y}^2_{h+1}, \ldots, \mathbf{Y}^t_{h+1}.$$

Proof. Define the following sets for $j \in [h]$:

$$\text{Ind}_j = \{i \in [2, h] : id^i[j] \neq id^1[j]\}, \quad \text{Ind}_{\leq j} = \cup_{g=1}^{j} \text{Ind}_g, \quad \overline{\text{Ind}_{\leq j}} = [t] \setminus \text{Ind}_{\leq j}.$$

We prove the following lemma from which Theorem 5.1 is direct by observing that $\text{Ind}_{\leq h} = [2, t]$.

Lemma 5.2. *For each $j \in [h]$,*

$$\mathbf{Y}^1_{j+1}, \{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\} \approx_{O(j\epsilon)} \mathbf{U}_{n_2}, \{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\}.$$

Proof. Recall that $\mathbf{R}_{c,j} = \text{LExt}(\mathbf{X} + \mathbf{Z}, \mathbf{S}_{c,j})$ (for any $c \in \{0, 1\}$ and $j \in [h]$). Define $\mathbf{R}_{c,j,\mathbf{X}} = \text{LExt}(\mathbf{X}, \mathbf{S}_{c,j})$ and $\mathbf{R}_{c,j,\mathbf{Z}} = \text{LExt}(\mathbf{Z}, \mathbf{S}_{c,j})$. Since LExt is linear seeded, it follows that $\mathbf{R}_{c,j} = \mathbf{R}_{c,j,\mathbf{X}} + \mathbf{R}_{c,j,\mathbf{Z}}$. Similarly, define $\overline{\mathbf{R}_{c,j,\mathbf{X}}} = \text{LExt}(\mathbf{X}, \overline{\mathbf{S}_{c,j}})$ and $\overline{\mathbf{R}_{c,j,\mathbf{Z}}} = \text{LExt}(\mathbf{Z}, \overline{\mathbf{S}_{c,j}})$.

We prove the lemma by induction on j. In fact, we prove the following stronger statement:

For every $j \in [0, h]$, conditioned on the random variables: $\{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\}, \{\mathbf{Y}^i_g : g \in [j], i \in [t]\}, \{\mathbf{R}^i_{0,j+1,\mathbf{Z}} : i \in \text{Ind}_j\}, \{\overline{\mathbf{Y}}^i_g : g \in [j], i \in [t]\}, \{\mathbf{S}^i_{0,g} : g \in [j], i \in [t]\}, \{\mathbf{S}^i_{1,g} : g \in [j], i \in [t]\}, \{\mathbf{R}^i_{0,g} : g \in [j], i \in [t]\}, \{\mathbf{R}^i_{1,g} : g \in [j], i \in [t]\}, \{\overline{\mathbf{S}^i_{0,g}} : g \in [j], i \in [t]\}, \{\overline{\mathbf{S}^i_{1,g}} : g \in [j], i \in [t]\}, \{\overline{\mathbf{R}^i_{0,g}} : g \in [j], i \in [t]\}, \{\overline{\mathbf{R}^i_{1,g}} : g \in [j], i \in [t]\}$

- $\mathbf{Y}^1_{j+1}$ is $6j\epsilon$-close to $\mathbf{U}_{n_2}$ on average
- $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- $\{\mathbf{Y}^i_{j+1} : i \in [t]\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.

- $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} = k+8td(h-j)+\log(1/\epsilon)$ and $\mathbf{Y}^1$ has conditional min-entropy at least $k_{j,\mathbf{Y}} = k + 10td(h-j) + 4tn_2(h-j+1) + \log(1/\epsilon)$.

The base case of the induction when $j = 0$ is direct. Now suppose the above holds for some $j-1 \geq 0$, and we prove it for j.

We fix the following random variables: $\{\mathbf{Y}^i_j : i \in \mathrm{Ind}_{\leq(j-1)}\}, \{\mathbf{Y}^i_g : g \in [j-1], i \in [t]\}$, $\{\overline{\mathbf{Y}}^i_g : g \in [j-1], i \in [t]\}, \{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in \mathrm{Ind}_{j-1}\}, \{\mathbf{S}^i_{0,g} : g \in [j-1], i \in [t]\}, \{\mathbf{S}^i_{1,g} : g \in [j-1], i \in [t]\}$, $\{\mathbf{R}^i_{0,g} : g \in [j-1], i \in [t]\}, \{\mathbf{R}^i_{1,g} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{S}^i_{0,g}} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{S}^i_{1,g}} : g \in [j-1]$, $i \in [t]\}, \{\overline{\mathbf{R}^i_{0,g}} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{R}^i_{1,g}} : g \in [j-1], i \in [t]\}$. By induction hypothesis, we have

- $\mathbf{Y}^1_j$ is $6(j-1)\epsilon$-close to $\mathbf{U}_{n_2}$ on average.
- $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- $\{\mathbf{Y}^i_j : i \in [t]\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- $\mathbf{X}$ has conditional min-entropy at least $k_{j-1,\mathbf{X}} = k_{j,\mathbf{X}} + 8td$ and $\mathbf{Y}^1$ has conditional min-entropy at least $k_{j-1,\mathbf{Y}} = k_{j,\mathbf{Y}} + 10td + 4tn_2$.

We repeatedly use Lemma 2.12 when we argue about the remaining conditional min-entropy in a random variable and do not explicitly mention this. Further, any random variable that we fix is either a deterministic function of $\mathbf{X}$ or a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Thus, we always maintain that $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ and again do not explicitly mention this.

We split the proof into two cases depending on the bit $id^1[j]$.

Case 1: Suppose $id^1[j] = 1$ and hence $\overline{\mathbf{Y}}^1_j = \mathrm{LExt}_3(\mathbf{Y}^1, \mathbf{R}^1_{1,j})$. It follows that for all $i \in \mathrm{Ind}_j$, $id^i[j] = 0$ and $\overline{\mathbf{Y}}^i_j = \mathrm{LExt}_3(\mathbf{Y}^i, \mathbf{R}^i_{0,j})$. Since $\{\mathbf{Y}^i_j : i \in \mathrm{Ind}_{\leq(j-1)}\}$ is fixed, it follows that for all $i \in \mathrm{Ind}_{\leq(j-1)}$, $\mathbf{R}^i_{0,j,\mathbf{X}} = \mathrm{LExt}_1(\mathbf{X}, \mathbf{S}^i_{0,j})$ is a deterministic function of $\mathbf{X}$. We fix the random variables $\{\mathbf{R}^i_{0,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, and $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 7td$. We now fix $\mathbf{S}^1_{0,j}$, $\{\mathbf{S}^i_{0,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in [t]\}$ and by Lemma 4.1, it follows that (a) $\mathbf{R}^1_{0,j}$ is $(6j-5)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 7td$ and $\mathbf{Y}^1_j$ has conditional min-entropy at least $k + td + \log(1/\epsilon)$. We also note that for each $i \in \mathrm{Ind}_{\leq(j-1)}$, $\mathbf{R}^i_{0,j} = \mathbf{R}^i_{0,j,\mathbf{X}} + \mathbf{R}^i_{0,j,\mathbf{Z}}$ is fixed.

Next we fix $\{\mathbf{S}^i_{1,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, observing that it is now a deterministic function of $\{\mathbf{Y}^i : i \in [t]\}$ and hence does not affect the distribution of $\mathbf{R}^1_{0,j}$. The conditional min-entropy of $\mathbf{Y}^1_j$ after this fixing is at least $k + \log(1/\epsilon)$. We now fix $\mathbf{R}^1_{0,j}$, $\{\mathbf{R}^i_{0,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ and by Lemma 4.1, (a) $\mathbf{S}^1_{1,j}$ is $(6j-4)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{Y}^1$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 6td$ and $\mathbf{Y}^1_j$ has conditional min-entropy at least $k + \log(1/\epsilon)$.

Continuing in a similar fashion as above, we first fix $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, which is a deterministic function of $\mathbf{X}$. The conditional min-entropy of $\mathbf{X}$ after this fixing is at least $k_{j,\mathbf{X}} + 5td$. We now fix the random variables $\mathbf{S}^1_{1,j}, \{\mathbf{S}^i_{1,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\mathbf{R}^i_{1,j,\mathbf{Z}} : i \in [t]\}$, $\{\mathbf{Y}^i_j : i \in [t]\}$ and by Lemma 4.1, we have (a) $\mathbf{R}^1_{1,j}$ is $(6j-3)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 5td$.

We fix $\{\overline{\mathbf{Y}}^i_j : i \in \mathrm{Ind}_{\leq(j-1)}\}$ which is deterministic function of $\{\mathbf{Y}^i : i \in [t]\}$, and $\mathbf{R}^1_{1,j}$ continues to remain close to $\mathbf{U}_d$ on average. We also fix $\{\overline{\mathbf{Y}}^i_j : i \in \mathrm{Ind}_j\}$ observing that it is a deterministic

function of $\{\mathbf{Y}^i : i \in [t]\}$ (since we have fixed $\{\mathbf{R}^i_{0,j} : i \in [t]\}$ and for $i \in \text{Ind}_j$, $\mathbf{Y}^i_j = \text{LExt}_3(\mathbf{Y}^i, \mathbf{R}^i_{0,j})$). It follows that $\{\overline{\mathbf{S}}^i_{0,j} : i \in \text{Ind}_{\leq j}\}$ is fixed and hence $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \text{Ind}_{\leq j}\}$ is a deterministic function of $\mathbf{Z}$. Thus, we fix $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \text{Ind}_{\leq j}\}$ without affecting the distribution of $\mathbf{R}^1_{1,j}$.

The conditional min-entropy of $\mathbf{Y}^1$ after this fixing is at least $k_{j,\mathbf{Y}} + 2tn_2 + 4td$. Thus $\overline{\mathbf{Y}}^1_j = \text{LExt}_3(\mathbf{Y}^1, \mathbf{R}^1_{1,j})$ is $(6j-2)\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\mathbf{R}^1_{1,j}$. We fix $\mathbf{R}^1_{1,j}$ and thus $\overline{\mathbf{Y}}^1_j$ is now a deterministic function of $\mathbf{Y}^1$. We now fix $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \overline{\text{Ind}_j}\}$ which is a deterministic function of $\mathbf{X}$ and note that this fixes $\{\mathbf{R}^i_{1,j} : j \in \overline{\text{Ind}_j}\}$. Further, since $\{\overline{\mathbf{Y}}^i_j : i \in \text{Ind}_{\leq j}\}$ is fixed, it follows that for all $i \in \text{Ind}_{\leq j}$, $\overline{\mathbf{R}}^i_{0,j,\mathbf{X}}$ is a deterministic function of $\mathbf{X}$. We fix the random variables $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{X}} : i \in \text{Ind}_{\leq j}\}$ and note that $\{\overline{\mathbf{S}}^i_{1,j} : i \in \text{Ind}_{\leq j}\}$ is now fixed. Thus $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \text{Ind}_{\leq j}\}$ is now a deterministic of $\mathbf{X}$. We fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \text{Ind}_{\leq j}\}$ and $\overline{\mathbf{Y}}^1_j$ continues to remain close to uniform on average and $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 2td$.

We now fix $\overline{\mathbf{S}}^1_{0,j}$, $\{\overline{\mathbf{S}}^i_{0,j} : i \in \overline{\text{Ind}_{\leq j}}\}$, $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \overline{\text{Ind}_{\leq j}}\}$, $\{\overline{\mathbf{Y}}^i_j : i \in \overline{\text{Ind}_{\leq j}}\}$ and by Lemma 4.1, it follows that (a) $\overline{\mathbf{R}}^1_{0,j}$ is $(6j-1)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 3td$. Next we fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{Z}} : i \in \text{Ind}_{\leq j}\}$ which a deterministic function of $\mathbf{Z}$ and $\{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\}$ is now a deterministic function of $\{\mathbf{Y}^i : i \in \text{Ind}_{\leq j}\}$. Thus, we fix $\{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\}$ and $\overline{\mathbf{R}}^1_{0,j}$ continues to remain uniform on average. It now follows that $\{\mathbf{R}^i_{0,j+1,\mathbf{Z}} : i \in \text{Ind}_{\leq (j)}\}$ is a deterministic function of $\mathbf{Z}$, and we fix it.

The conditional min-entropy of $\mathbf{Y}^1$ after this fixing is at least $k_{j,\mathbf{Y}}$ and thus, $\mathbf{Y}^1_{j+1} = \text{LExt}_3(\mathbf{Y}^1, \overline{\mathbf{R}}^1_{0,j})$ is $6j\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\overline{\mathbf{R}}^1_{0,j}$. We fix $\overline{\mathbf{R}}^1_{0,j}$ which is a deterministic function of $\mathbf{X}$ and thus $\mathbf{Y}^1_{j+1}$ is now a deterministic function of $\mathbf{Y}^1$. Now consider any $i \in \overline{\text{Ind}_{\leq j}}$. Since we have fixed $\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}}$ and $\overline{\mathbf{R}}^i_{0,j} = \overline{\mathbf{R}}^i_{0,j,\mathbf{X}} + \overline{\mathbf{R}}^i_{0,j,\mathbf{Z}}$, it follows that $\overline{\mathbf{R}}^i_{0,j,\mathbf{X}}$ is a deterministic function of $\mathbf{X}$. Thus, we fix $\{\overline{\mathbf{R}}^i_{0,j} : i \in \overline{\text{Ind}_{\leq j}}\}$ without affecting the distribution of $\mathbf{Y}^1_{j+1}$. $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + td$ after this fixing. Now, since $\overline{\mathbf{Y}}^i_j$ is fixed, it follows that $\overline{\mathbf{S}}^1_{1,j}$ is fixed for each $i \in [t]$. Thus, for any $i \in \overline{\text{Ind}_{\leq j}}$, $\overline{\mathbf{R}}^i_{1,j,X} = \text{LExt}_1(\mathbf{X}, \overline{\mathbf{S}}^i_{1,j})$ is a deterministic function of $\mathbf{X}$. We fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \overline{\text{Ind}_{\leq j}}\}$, and observe that $\mathbf{Y}^1_{j+1}$ remains close to uniform on average and $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}}$. Thus, $\{\overline{\mathbf{R}}^i_{1,j} : i \in \overline{\text{Ind}_{\leq j}}\}$ is now a deterministic function of $\mathbf{Z}$ and $\{\mathbf{Y}^i_{j+1} : i \in \overline{\text{Ind}_{\leq j}}\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. This concludes the proof of this case.

Case 2: Suppose $id^1[j] = 0$ and hence $\overline{\mathbf{Y}}^1_j = \text{LExt}_3(\mathbf{Y}^1, \mathbf{R}^1_{0,j})$. Since $\{\mathbf{Y}^i_j : j \in \text{Ind}_{j-1}\}$ is fixed, it follows that $\{\mathbf{R}^i_{0,j,\mathbf{X}} : i \in \text{Ind}_{\leq (j-1)}\}$ and $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \text{Ind}_{\leq (j-1)}\}$ are deterministic functions of $\mathbf{X}$ and we fix them without affecting the distribution of $\mathbf{Y}^1_j$. $\mathbf{X}$ has conditional min-entropy at least $k_{j-1,\mathbf{X}} + 6td$ after this fixing.

We now fix $\mathbf{S}^1_{0,j}, \{\mathbf{S}^i_{0,j} : i \in \overline{\text{Ind}_{\leq (j-1)}}\}$, $\mathbf{R}^1_{0,j,\mathbf{Z}}$, $\{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in \overline{\text{Ind}_{\leq (j-1)}}\}$ and by Lemma 4.1, $\mathbf{R}^1_{0,j}$ is $(6j-5)\epsilon$-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{X}$. We next fix $\{\mathbf{R}^i_{1,j,\mathbf{Z}} : i \in \text{Ind}_{\leq (j-1)}\}$, $\{\mathbf{Y}^i_j : i \in \overline{\text{Ind}_{\leq (j-1)}}\}$, and $\{\overline{\mathbf{Y}}^i_j : i \in \text{Ind}_{\leq (j-1)}\}$ observing that they are deterministic functions of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ and does not affect the distribution of $\mathbf{R}^1_{0,j}$. Further, $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \text{Ind}_{\leq (j-1)}\}$ is now a deterministic function of $\mathbf{Z}$, and we fix it.

As a result of these fixings, $\mathbf{Y}^1$ has conditional min-entropy at least $k_{j-1,\mathbf{Y}} + 5tdh + 2tn_2$. Thus, $\overline{\mathbf{Y}}^1_j$ is $(6j-4)\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\mathbf{R}^1_{0,j}$. We fix $\mathbf{R}^1_{0,j}$ and $\overline{\mathbf{Y}}^1_j$ is now a

deterministic function of $\mathbf{Y}^1$. We now fix $\{\mathbf{R}^i_{0,j,X} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ which is a deterministic function of $\mathbf{X}$ and note that this fixes $\{\mathbf{S}^i_{0,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$. Thus $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$ is now a deterministic function of $\mathbf{X}$ and we fix it without affecting the distribution of $\overline{\mathbf{Y}}^i_j$. As a result of this fixing $\{\mathbf{R}^i_{1,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ is a deterministic function of $\mathbf{Z}$ and hence $\{\overline{\mathbf{Y}}^i_j : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Next, we fix $\{\overline{\mathbf{R}}^i_{0,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$ and $\{\overline{\mathbf{R}}^i_{1,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, noting that they are deterministic functions of $\mathbf{X}$. $\mathbf{X}$ has conditional min-entropy at least $k_{j-1,\mathbf{X}} + 2td$ after these fixings.

We now fix $\overline{\mathbf{S}}^1_{0,j}$, $\{\overline{\mathbf{S}}^i_{0,j} : i \in \overline{\mathrm{Ind}_{\leq(h-1)}}\}$, $\overline{\mathbf{R}}^1_{0,j,\mathbf{Z}}$, $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} i \in \overline{\mathrm{Ind}_{\leq(h-1)}}\}$ and invoking Lemma 4.1, it follows that $\overline{\mathbf{R}}^1_{0,j}$ is $(6j-3)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$. We now fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{Z}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$ which a deterministic function of $\mathbf{Z}$ and note that this fixes $\{\overline{\mathbf{R}}^i_{1,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$. Further $\overline{\mathbf{Y}}^1_j$ has conditional min-entropy at least $k + td + \log(1/\epsilon)$. We now fix $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\overline{\mathbf{S}}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, and by Lemma 4.1, it follows that $\overline{\mathbf{R}}^i_{1,j}$ is $(6j-1)\epsilon$-close to $\mathbf{U}_d$ on average and is deterministic function of $\mathbf{X}$.

We now observe that $\{\mathbf{Y}^i_{j+1} : i \in \mathrm{Ind}_{\leq j}\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ and fix it without affecting the distribution of $\overline{\mathbf{R}}^1_{1,j}$. Next we fix $\{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in \mathrm{Ind}_{\leq j}\}$ which is now a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. The conditional min-entropy of $\mathbf{Y}^1$ is at least $k_{j,\mathbf{Y}}$ and hence $\mathbf{Y}^i_{j+1}$ is $6j\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\overline{\mathbf{R}}^1_{1,j}$. We fix $\overline{\mathbf{R}}^1_{1,j}$ and thus $\mathbf{Y}^1_{j+1}$ is now a deterministic function of $\mathbf{Y}^1$. Thus we fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq j}}\}$ and as a result $\{Y^i_{j+1} : i \in \overline{\mathrm{Ind}_{\leq j}}\}$ is now a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Further $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}}$ as a result of these fixings. This completes the proof of induction and the theorem follows. □

□

6 Extractors for Sumset Sources

In this section we construct explicit extractors for (n, k, C)-sumset sources where $k = \mathrm{polylog}(n)$ and C is a large enough constant.

Theorem 6.1 (Theorem 1 restated)**.** *There exists constants $c, C > 0$ and a small constant $\beta_1 > 0$ such that for all $n \in \mathbb{N}$, there exists a polynomial time computable extractor for $(n, k, C+1)$-sumset sources, $k \geq \log^c(n)$, with error $n^{-\Omega(1)}$ and output length k^{β_1}.*

We use the rest of the section to prove Theorem 6.1. We claim that the function computed by Algorithm 3 is the required extractor. We first set up the parameters and ingredients used by Algorithm 3.

- Let $\beta = 1/20$, $t = k^\beta, \epsilon = 1/n^2$.
- Let $c = (\lambda + 1)/\beta$.
- Let $\mathrm{LExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{n_1}$, $n_1 = \sqrt{k}$, be the linear seeded extractor from Theorem 2.3 set extract from min-entropy k with error ϵ. Thus $d = c_1 \log n$, for some constant c_1. Let $D = 2^d = n^{c_1}$.
- Let $C = c_1 + 2, k' = d^2, \epsilon_1 = 1/D^{2t} = 1/n^{2tc_1}, n_2 = k^{4\beta}, k'' = n_2^2 = k^{8\beta}, \delta = (2c_1 - 1)/2c_1$.

- Let $\text{LExt}_1 : \{0,1\}^{n_1} \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$ and $\text{LExt}_2 : \{0,1\}^{n_2} \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$ be instantiations of the linear seeded extractor from Theorem 2.2, both set to extract from min-entropy k' with error ϵ_1. Thus, $d_1 = O(\log^2(k/\epsilon_1)) = O(t^2 \log^2 n)$ and $d_2 = O(\log^2(k/\epsilon_1)) = O(t^2 \log^2 n)$. Finally let $\text{LExt}_3 : \{0,1\}^{n_1} \times \{0,1\}^{d_3} \to \{0,1\}^{n_2}$ be an instantiation of the linear seeded extractor from Theorem 2.2 set to extract from min-entropy k'' with error ϵ_1. Thus, $d_3 = O(\log^2(n_1/\epsilon_1)) = O(t^2 \log^2 n)$. Let LCB be the function computed by Algorithm 2 using these linear seeded extractors.
- Let $\text{bitExt} : \{0,1\}^D \to \{0,1\}^m$, $m = t^\alpha$, be the extractor from Theorem 2.7 set to extract from (q, t, γ)-non-oblivious sources where $q = D^\delta$ and $\gamma = 1/D^{t+1}$.

Algorithm 3: SUMExt(x)

Input: A bit string $x = x_1 + \ldots + x_{C+1}$, where each x_i is a bit string of length n.
Output: A bit string of length m.

1 Let w be the $n_1 \times D$ boolean matrix whose i^{th} row w_i is given by $\text{LExt}(x, s_i)$.
2 Let v be the $n_2 \times D$ boolean matrix whose i^{th} row v_i is given by $\text{LCB}(w_i, x, s_i)$.
3 Let r be the first column of the matrix v. Output $\text{bitExt}(r)$.

We prove the following claims about the random variables computed in Algorithm 3 from which Theorem 6.1 is direct.

Claim 6.2. $\mathbf{V}$ *is* $1/n^{O(1)}$*-close to a somewhere-random source* $\mathbf{V}'$ *containing a subset* R *of rows,* $|R| \geq D - D^\delta$ *such that the joint distribution of any* t *distinct rows in* R *is* γ*-close to* $\mathbf{U}_{tm}$*.*

Proof. Since LExt is a strong seeded extractor, it follows that for any $j \in [C]$, there exists a subset $S_j \subset \{0,1\}^d$, $|S_j| \geq (1 - \sqrt{\epsilon})D$, such that for any $s \in S_j$ $\text{LExt}(\mathbf{X}, s_j)$ is $\sqrt{\epsilon}$-close to $\mathbf{U}_{n_1}$. Thus, by a union bound, it follows that there exists a set $S \subset \{0,1\}^d$,

$$|S| \geq (1 - C\sqrt{\epsilon})D > D - D^\delta,$$

(the inequality follows by our choice of parameters) such that for any $s_i \in S$, $\text{LExt}(\mathbf{X}_j, s_i)$ is $\sqrt{\epsilon}$-close to $\mathbf{U}_{n_1}$ for each $j \in [C]$.

Since LExt is linear seeded, it follows that for any $i \in [D]$, it follows that $\mathbf{W}^i = \text{LExt}(\mathbf{X}, s_i) = \left(\sum_{j=1}^{C} \text{LExt}(\mathbf{X}_j, s_i)\right) + \text{LExt}(\mathbf{X}_{C+1}, s_i)$. Thus if $s_i \in S$, then by Lemma 2.14, $\left(\sum_{j=1}^{C} \text{LExt}(\mathbf{X}_j, s_i)\right)$ is $\epsilon^{C/2}$-close to $\mathbf{U}_{n_1}$. Using a hybrid argument, it follows that $\mathbf{W}$ is $D\epsilon^{C/2}$-close to a $D \times n_1$ matrix $\overline{\mathbf{W}}$, whose i^{th} row $\overline{\mathbf{W}}^i$ is equal to $\mathbf{W}^i$ if $s_i \notin S$, and otherwise is given by $\mathbf{Y}^i + \text{LExt}(\mathbf{X}_{C+1}, s_i)$, where $\mathbf{Y}^i$ follows the distribution $\mathbf{U}_{n_2}$. We note that the $\mathbf{Y}^i$'s can be arbitrarily correlated.

Thus, $\mathbf{V}$ is $D\epsilon^{C/2}$-close to a $D \times n_2$-matrix $\overline{\mathbf{V}}$ such that if $s_i \in S$, then the i^{th} row $\overline{\mathbf{V}}^i$ is given by $\text{LCB}(\mathbf{Y}^i + \text{LExt}(\mathbf{X}_{C+1}, s_i), \mathbf{X}, s_i)$.

Now consider any subset $\{s_{i_1}, \ldots, s_{i_t}\} \subset S$ of size t. We claim that

$$(\overline{\mathbf{V}}^{i_1}, \ldots, \overline{\mathbf{V}}^{i_t}) \approx_{O(td\epsilon)} \mathbf{U}_{tm}.$$

We fix the random variable $\{\text{LExt}(\mathbf{X}_{C+1}, s_{i_1}), \ldots, \text{LExt}(\mathbf{X}_{C+1}, s_{i_t})\}$. As a result of this fixing, $\mathbf{X}_{C+1}$ has min-entropy at least $k - tn_1 - \log(1/\epsilon) > k/2$ with probability at least $1 - \epsilon$. Let $\mathbf{Z} = \sum_{j=1}^{C} \mathbf{X}_j$.

Thus,

$$(\overline{\mathbf{V}}^{i_1}, \ldots, \overline{\mathbf{V}}^{i_t}) = (\mathrm{LCB}(\mathbf{Y}^1 + a_1, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_1}), \ldots, \mathrm{LCB}(\mathbf{Y}^t + a_t, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_t})),$$

where $a_1, \ldots, a_t$ are some constants.

We now invoke Theorem 5.1 noting that the following conditions hold by our choice of parameters:

- $\mathbf{X}_{C+1}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- Each s_{i_g} is a distinct bit string of length d.
- $k/2 \geq k' + 8td_1d + \log(1/\epsilon)$.
- $n_2 \geq k' + 3td_1 + \log(1/\epsilon)$.
- $n_1 \geq k' + 10td_1d + (4td + 1)n_2 + \log(1/\epsilon)$.

Thus,

$$(\mathrm{LCB}(\mathbf{Y}^1 + a_1, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_1}), \ldots, \mathrm{LCB}(\mathbf{Y}^t + a_t, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_t})) \approx_{O(dt\epsilon_1)} \mathbf{U}_{tm}.$$

We note that by our choice of parameters, the following inequalities hold:

- $dt\epsilon_1 < 1/D^{t+2}$.
- $\epsilon^{C/2}D \leq 1/n^2$.

The claim now follows from the fact the above argument holds for any arbitrary size t subset of S and the fact that $\mathbf{V}$ is $\epsilon^{C/2}D$-close to $\overline{\mathbf{V}}$. □

Claim 6.3. *$\mathbf{V}'$ is $1/n^{O(1)}$-close to $\mathbf{U}_m$.*

Proof. Follows directly from Claim 6.2 and Theorem 2.7. □

7 Proof of Theorem 1.12

Proof of Theorem 1.12. Let $\mathrm{2Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$, $m = k/10$ be the 2-source extractor from Theorem 2.4 set to extract from min-entropy $k/2$ with error $\epsilon = 1/n^{\Omega(1)}$. Define the function $\mathrm{Ext} : \{0,1\}^{\ell n} \to \{0,1\}^m$ as

$$\mathrm{Ext}(x_1, \ldots, x_\ell) = \sum_{1 \leq i < j \leq \ell} \mathrm{2Ext}(x_i, x_j).$$

We claim that for any (n, k, ℓ)-somewhere 2-source $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_\ell\}$,

$$|\mathrm{Ext}(\mathbf{X}) - \mathbf{U}_m| \leq \epsilon.$$

We prove this in the following way. Since the function Ext is symmetric, we can assume without loss of generality that the sources $\mathbf{X}_1$ and $\mathbf{X}_2$ have min-entropy at least k each. Fix the sources $\mathbf{X}_3, \ldots, \mathbf{X}_\ell$. Thus, after this fixing

$$\mathrm{Ext}(\mathbf{X}_1, \mathbf{X}_2, x_3 \ldots, x_\ell) = \mathrm{2Ext}(\mathbf{X}_1, \mathbf{X}_2) + \left(\sum_{j=3}^{\ell} \mathrm{2Ext}(\mathbf{X}_1, x_j)\right) + \left(\sum_{j=3}^{\ell} \mathrm{2Ext}(\mathbf{X}_2, x_j)\right) + s,$$

for some constant $s \in \{0,1\}^m$. Now, we observe that $\mathbf{A} = \left(\sum_{j=3}^{\ell} 2\mathrm{Ext}(\mathbf{X}_1, x_j)\right)$ is a random variable on $\{0,1\}^m$ and is deterministic function of $\mathbf{X}_1$. Thus, we fix $\mathbf{A}$, and using Lemma 2.13, $\mathbf{X}_1$ has min-entropy at least $0.9k - m$ with probability $1 - 2^{-k^{0.1}}$. Similarly, $\mathbf{B} = \left(\sum_{j=3}^{\ell} 2\mathrm{Ext}(\mathbf{X}_2, x_j)\right)$ is a random variable on $\{0,1\}^m$ and is deterministic function of $\mathbf{X}_2$. Thus, we fix $\mathbf{B}$, and $\mathbf{X}_2$ has min-entropy at least $0.9k - m$ with probability $1 - 2^{-k^{0.1}}$. Thus, after this fixing

$$\mathrm{Ext}(\mathbf{X}) = 2\mathrm{Ext}(\mathbf{X}_1, \mathbf{X}_2) + s',$$

for some constant $s' \in \{0,1\}^m$. Further $\mathbf{X}_1$ and $\mathbf{X}_2$ are still independent, each with min-entropy at least $0.8k$ (with probability at least $1 - 2^{-k^{\Omega(1)}}$). The result now follows since 2Ext is a 2-source extractor for min-entropy $k/2$. □

Acknowledgments

We thank David Zuckerman for his collaboration during the initial stages of this paper.

References

[BIW06] Boaz Barak, Russell Impagliazzo, and Avi Wigderson. Extracting randomness using few independent sources. *SIAM J. Comput.*, 36(4):1095–1118, December 2006.

[BK12] Eli Ben-Sasson and Swastik Kopparty. Affine dispersers from subspace polynomials. *SIAM J. Comput.*, 41(4):880–914, 2012.

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

[Blu86] Manuel Blum. Independent unbiased coin flips from a correlated biased sourcea finite state markov chain. *Combinatorica*, 6(2):97–108, 1986.

[Bou05] J. Bourgain. More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 01(01):1–32, 2005.

[Bou07] Jean Bourgain. On the construction of affine extractors. *GAFA Geometric And Functional Analysis*, 17(1):33–57, 2007.

[BRSW12] Boaz Barak, Anup Rao, Ronen Shaltiel, and Avi Wigderson. 2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction. *Annals of Mathematics*, 176(3):1483–1543, 2012. Preliminary version in STOC '06.

[BSZ11] Eli Ben-Sasson and Noga Zewi. From affine to two-source extractors via approximate duality. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, 2011.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[CGH+85] Benny Chor, Oded Goldreich, Johan Hastad, Joel Friedman, Steven Rudich, and Roman Smolensky. The bit extraction problem of t-resilient functions (preliminary version). In *26th Annual Symposium on Foundations of Computer Science, Portland, Oregon, USA, 21-23 October 1985*, pages 396–407, 1985.

[CGL15] Eshan Chattopadhyay, Vipul Goyal, and Xin Li. Non-malleable extractors and codes, with their many tampered extensions. *CoRR*, abs/1505.00107, 2015.

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

[CZ15a] Eshan Chattopadhyay and David Zuckerman. Explicit two-source extractors and resilient functions. Technical Report TR15-119, ECCC, 2015.

[CZ15b] Eshan Chattopadhyay and David Zuckerman. New extractors for interleaved sources. Technical Report TR15-151, ECCC, 2015.

[DK11] Evgeny Demenkov and Alexander S. Kulikov. An elementary proof of a 3n - o(n) lower bound on the circuit complexity of affine dispersers. In *Proceedings of the 36th International Conference on Mathematical Foundations of Computer Science*, MFCS'11, pages 256–265, Berlin, Heidelberg, 2011. Springer-Verlag.

[DORS08] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. *SIAM Journal on Computing*, 38:97–139, 2008.

[DP07] Stefan Dziembowski and Krzysztof Pietrzak. Intrusion-resilient secret sharing. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 227–237, Washington, DC, USA, 2007. IEEE Computer Society.

[DW09] Yevgeniy Dodis and Daniel Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In *STOC*, pages 601–610, 2009.

[FGHK15] Magnus Gausdal Find, Alexander Golovnev, Edward Hirsch, and Alexander Kulikov. A better-than-3n lower bound for the circuit complexity of an explicit function. Technical Report TR15-166, ECCC, 2015.

[GR08] Ariel Gabizon and Ran Raz. Deterministic extractors for affine sources over large fields. *Combinatorica*, 28(4):415–440, 2008.

[GRS06] Ariel Gabizon, Ran Raz, and Ronen Shaltiel. Deterministic extractors for bit-fixing sources by obtaining an independent seed. *SIAM J. Comput.*, 36(4):1072–1094, 2006.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil P. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *J. ACM*, 56(4), 2009.

[KLR09] Yael Kalai, Xin Li, and Anup Rao. 2-source extractors under computational assumptions and cryptography with defective randomness. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 617–628, 2009.

[KLRZ08] Yael Tauman Kalai, Xin Li, Anup Rao, and David Zuckerman. Network extractor protocols. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 654–663, 2008.

[KRVZ11] Jesse Kamp, Anup Rao, Salil P. Vadhan, and David Zuckerman. Deterministic extractors for small-space sources. *Journal of Computer and System Sciences*, 77:191–220, 2011.

[KZ07] Jesse Kamp and David Zuckerman. Deterministic Extractors for Bit-Fixing Sources and Exposure-Resilient Cryptography. *Siam Journal on Computing*, 36:1231–1247, 2007.

[Li11] Xin Li. A new approach to affine extractors and dispersers. In *Computational Complexity (CCC), 2011 IEEE 26th Annual Conference on*, pages 137–147, June 2011.

[Li13a] Xin 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*, pages 100–109, 2013.

[Li13b] Xin Li. New independent source extractors with exponential improvement. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pages 783–792, 2013.

[Li15a] Xin Li. Extractors for affine sources with polylogarithmic entropy. Technical Report TR15-121, ECCC, 2015.

[Li15b] Xin Li. Improved constructions of two-source extractors. Technical Report TR15-125, ECCC, 2015.

[Li15c] Xin Li. Three-source extractors for polylogarithmic min-entropy. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, 2015.

[MW97] Ueli Maurer and Stefan Wolf. Privacy amplification secure against active adversaries. In *Advances in Cryptology — CRYPTO '97*, volume 1294, pages 307–321, August 1997.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *J. Comput. Syst. Sci.*, 52(1):43–52, 1996.

[Rao09a] Anup Rao. Extractors for a constant number of polynomially small min-entropy independent sources. *SIAM J. Comput.*, 39(1):168–194, 2009.

[Rao09b] Anup Rao. Extractors for low-weight affine sources. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity*, 2009.

[Raz05] Ran Raz. Extractors with weak random seeds. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 11–20, 2005.

[RRV02] Ran Raz, Omer Reingold, and Salil Vadhan. Extracting all the randomness and reducing the error in Trevisan's extractors. *JCSS*, 65(1):97–128, 2002.

[RY11] Ran Raz and Amir Yehudayoff. Multilinear formulas, maximal-partition discrepancy and mixed-sources extractors. *Journal of Computer and System Sciences*, 77:167–190, 2011.

[Sha08] Ronen Shaltiel. How to get more mileage from randomness extractors. *Random Struct. Algorithms*, 33(2):157–186, 2008.

[Sha11] Ronen Shaltiel. Dispersers for affine sources with sub-polynomial entropy. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 247–256, 2011.

[Tre01] Luca Trevisan. Extractors and pseudorandom generators. *Journal of the ACM*, pages 860–879, 2001.

[TV00] Luca Trevisan and Salil P. Vadhan. Extracting Randomness from Samplable Distributions. In *IEEE Symposium on Foundations of Computer Science*, pages 32–42, 2000.

[Vio14] Emanuele Viola. Extractors for circuit sources. *SIAM J. Comput.*, 43(2):655–672, 2014.

[vN51] J. von Neumann. Various techniques used in connection with random digits. *Applied Math Series*, 12:36–38, 1951. Notes by G.E. Forsythe, National Bureau of Standards. Reprinted in *Von Neumann's Collected Works*, 5:768-770, 1963.

[Yeh11] Amir Yehudayoff. Affine extractors over prime fields. *Combinatorica*, 31(2):245–256, 2011.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de