Electronic Colloquium on Computational Complexity, Report No. 25 (2013)

# Extractors for a Constant Number of Independent Sources with Polylogarithmic Min-Entropy

Xin Li*
Department of Computer Science
University of Washington
Seattle, WA 98905, U.S.A.
lixints@cs.washington.edu

February 6, 2013

**Abstract**

We study the problem of constructing explicit extractors for independent general weak random sources. Given weak sources on $n$ bits, the probabilistic method shows that there exists a deterministic extractor for two independent sources with min-entropy as small as $\log n + O(1)$. However, even to extract from a constant number of independent sources, previously the best known extractors require the min-entropy to be at least $n^{\delta}$ for any constant $\delta > 0$ [Rao06, BRSW06, Li13]. For sources on $n$ bits with min-entropy $k$, previously the best known extractor needs to use $O(\log(\log n/\log k)) + O(1)$ independent sources [Li13].

In this paper, we construct the first explicit extractor for a constant number of independent sources on $n$ bits with min-entropy $k \geq \text{polylog}(n)$. Thus, for the first time we get extractors for independent sources that are close to optimal. Our extractor is obtained by improving the condenser for structured somewhere random sources in [Li13], which is based on a connection between the problem of condensing somewhere random sources and the problem of leader election in distributed computing.

*Supported by a Simons postdoctoral fellowship.

ISSN 1433-8092

# 1 Introduction

*Randomness extractors* are functions that transform defective random sources into nearly uniform distributions. The original motivation for such objects comes from the fact that randomness plays an important role in computation, while real world random sources rarely satisfy the requirements of these applications. Indeed, these applications (e.g., algorithms, distributed computing and cryptography) typically require the random bits to be uniform, while in the real world random sources are almost always biased. In addition, even original uniform random bits used in cryptographic applications can be compromised as a result of side channel attacks. Therefore, it is important to study how to run these applications with imperfect randomness. Here we model imperfect randomness as an arbitrary distribution with a certain amount of entropy, and we use the standard min-entropy to measure the randomness in a random source $X$.

**Definition 1.1.** The *min-entropy* of a random variable $X$ is

$$H_\infty(X) = \min_{x \in \mathsf{supp}(X)} \log_2(1/\Pr[X = x]).$$

For $X \in \{0,1\}^n$, we call $X$ an $(n, H_\infty(X))$-source, and we say $X$ has *entropy rate* $H_\infty(X)/n$.

Randomness extractors provide us a way to bridge the above gap. Ideally, one would like to construct a deterministic algorithm such that given any imperfect random source with a certain amount of entropy as the input, the algorithm outputs a distribution that is almost uniform. However, this is easily shown to be impossible. Given this negative result, the study of randomness extractors has been pursued in two different directions.

In [NZ96], Nisan and Zuckerman introduced the notion of *seeded extractors*. These extractors are given an additional independent truly uniform random string as the seed. Typically the length of the seed is much shorter than the length of the input source, say $d = O(\log n)$. With the help of the seed randomness extraction becomes possible. We note that seeded extractors are useful even in situations where uniform random bits are not available, for example simulating randomized algorithms using weak sources, just by trying all possible values of the seed. Seeded extractors are related to many other areas in computer science, and today the constructions of such extractors are nearly optimal [LRVW03, GUV09, DW08].

However, in many other applications such as distributed computing and cryptography, it is not clear how to use seeded extractors with the above trick. Instead, we need extractors that only use weak random sources as inputs. Since it is impossible to extract from a single weak random source, one natural direction is to try to build extractors for multiple independent weak random sources. After all, it does not seem much stronger to assume that we have several independent weak sources in nature than that we have one such source. These extractors are called independent source extractors. Formally, we have the following definition.

**Definition 1.2** (Independent Source Extractor)**.** A function $\mathsf{IExt} : (\{0,1\}^n)^t \to \{0,1\}^m$ is an extractor for independent $(n,k)$ sources that uses $t$ sources and outputs $m$ bits with error $\epsilon$, if for any $t$ independent $(n,k)$ sources $X_1, X_2, \cdots, X_t$, we have

$$|\mathsf{IExt}(X_1, X_2, \cdots, X_t) - U_m| \leq \epsilon,$$

where $|\cdot|$ denotes the statistical distance.

The study of independent source extractors has a long history. For example, the well known Lindsey's lemma gives an extractor for two independent $(n, k)$ sources with $k > n/2$. These extractors are indeed used in applications in distributed computing and cryptography, for example the network extractor protocols in [KLRZ08, KLR09]. To get a sense of what kind of parameters one can achieve, it is easy to show by the probabilistic method that a deterministic extractor exists for just two independent sources with logarithmic min-entropy. In fact, most random functions are very good two-source extractors. Thus, constructing explicit independent source extractors is also closely related to the general problem of *derandomization*.

Another important reason for studying such extractors is their close connection to Ramsey graphs. For example, any function with two $n$-bit inputs gives a bipartite graph with $N = 2^n$ vertices on each side, where two vertices are connected if and only if the first bit of the output is 1. In this view, a two-source extractor for $(n, k)$ sources gives a bipartite graph with $N = 2^n$ vertices on each side, such that there is no bipartite clique or independent set of size $K = 2^k$. The bipartite Ramsey graph can also be converted to a regular Ramsey graph. More generally, extractors that use a few (say a constant) number of sources give Ramsey hypergraphs.

However, despite considerable efforts on finding explicit independent source extractors, the known constructions are far from optimal, and the problem of constructing better independent source extractors is a major open problem in the area of *pseudorandomness*. Below we review some of the previous constructions.

The formal study of extractors for independent sources started with Chor and Goldreich [CG88], where they constructed explicit extractors for two independent $(n, k)$ sources with $k \geq (1/2+\delta)n$ for any constant $\delta > 0$. After that there had been essentially no progress until Barak, Impagliazzo and Wigderson [BIW04] showed how to extract from a constant number $(\text{poly}(1/\delta))$ of independent $(n, \delta n)$ sources, for any constant $\delta > 0$. Their work was based on new techniques developed in additive combinatorics, e.g., sum-product theorems and incidence theorems. Following this work, more involved constructions appeared. Barak et al. [BKS+05] constructed extractors for three independent $(n, \delta n)$ sources for any constant $\delta > 0$, and this was later improved by Raz [Raz05] to given an extractor that works for three independent sources where only one is required to be an $(n, \delta n)$ source while the other two can have entropy as small as $k \geq \text{polylog}(n)$. In the same paper Raz also gave an extractor for two independent sources where one is required to have entropy $k \geq (1/2 + \delta)n$ for any constant $\delta > 0$, and the other can have entropy as small as $k \geq \text{polylog}(n)$.

Using more techniques from additive combinatorics, Bourgain [Bou05] gave an extractor that works for two independent sources with entropy $k \geq (1/2 - \delta)n$ for some universal constant $\delta > 0$, and this is currently the best known two-source extractor.

By using ideas related to somewhere random sources, Rao [Rao06] and subsequently Barak et al. [BRSW06] constructed extractors for general $(n, k)$ sources that use $O(\log n/\log k)$ independent sources. Thus for any constant $\delta > 0$ and entropy $k \geq n^\delta$ their extractors use only a constant number of sources. Li [Li11] constructed extractors for three independent sources with entropy $k \geq n^{1/2+\delta}$ for any constant $\delta > 0$, and this is currently the best known three-source extractor.

Recently, Li [Li13] constructed a new extractor for independent $(n, k)$ sources that uses only $O(\log(\log n/\log k)) + O(1)$ sources. This improves the results of [Rao06, BRSW06] exponentially. However, similar as in [Rao06, BRSW06], to extract from a constant number of sources this extractor still needs the entropy of the source to be at least $n^\delta$ for some constant $\delta > 0$. Therefore, a natural open problem is to see if we can construct extractors for a constant number of independent sources with sub-polynomially small min-entropy.

### 1.1 Our results

In this paper, we significantly improve all previous results. In fact, we build extractors for independent sources with nearly optimal parameters. We construct extractors for a constant number of independent sources with poly-logarithmic min-entropy. Our main result is as follows.

**Theorem 1.3.** *For every constant $\eta > 0$ and all $n, k \in \mathbb{N}$ with $k \geq \log^{2+\eta} n$, there exists an explicit function $\mathsf{IExt} : (\{0,1\}^n)^t \to \{0,1\}^m$ with $m = \Omega(k)$ and $t = O\left(\frac{1}{\eta}\right) + O(1)$ such that if $(X_1, \cdots, X_t)$ are $t$ independent $(n,k)$ sources, then*

$$|\mathsf{IExt}(X_1, \cdots, X_t) - U_m| \leq \epsilon,$$

*where $\epsilon = 1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$.*

We also have the following two corollaries.

**Corollary 1.4.** *For all $n, k \in \mathbb{N}$ with $k \geq \log^3 n$, there is an explicit extractor that uses $O(1)$ independent $(n,k)$ sources and outputs $m = \Omega(k)$ bits with error $1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$.*

**Corollary 1.5.** *For every constant $0 < \eta < 1$ and all $n, k \in \mathbb{N}$ with $k \geq \log^{2+\eta} n$, there is an explicit extractor that uses $O\left(\frac{1}{\eta}\right)$ independent $(n,k)$ sources and outputs $m = \Omega(k)$ bits with error $1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$.*

Table 1 summarizes our results compared to previous constructions of independent source extractors.

| Construction | Number of Sources | Min-Entropy | Output | Error |
|---|---|---|---|---|
| [CG88] | 2 | $k \geq (1/2 + \delta)n$, any constant $\delta$ | $\Theta(n)$ | $2^{-\Omega(n)}$ |
| [BIW04] | $\mathrm{poly}(1/\delta)$ | $\delta n$, any constant $\delta$ | $\Theta(n)$ | $2^{-\Omega(n)}$ |
| [BKS+05] | 3 | $\delta n$, any constant $\delta$ | $\Theta(1)$ | $O(1)$ |
| [Raz05] | 3 | One source: $\delta n$, any constant $\delta$. Other sources may have $k \geq \mathrm{polylog}(n)$. | $\Theta(1)$ | $O(1)$ |
| [Raz05] | 2 | One source: $(1/2 + \delta)n$, any constant $\delta$. Other source may have $k \geq \mathrm{polylog}(n)$ | $\Theta(k)$ | $2^{-\Omega(k)}$ |
| [Bou05] | 2 | $(1/2 - \alpha_0)n$ for some small universal constant $\alpha_0 > 0$ | $\Theta(n)$ | $2^{-\Omega(n)}$ |
| [Rao06] | 3 | One source: $\delta n$, any constant $\delta$. Other sources may have $k \geq \mathrm{polylog}(n)$. | $\Theta(k)$ | $2^{-k^{\Omega(1)}}$ |
| [Rao06] | $O(\log n / \log k)$ | $k \geq \mathrm{polylog}(n)$ | $\Theta(k)$ | $k^{-\Omega(1)}$ |
| [BRSW06] | $O(\log n / \log k)$ | $k \geq \mathrm{polylog}(n)$ | $\Theta(k)$ | $2^{-k^{\Omega(1)}}$ |
| [Li11] | 3 | $k = n^{1/2+\delta}$, any constant $\delta$ | $\Theta(k)$ | $k^{-\Omega(1)}$ |
| [Li13] | $O(\log(\frac{\log n}{\log k})) + O(1)$ | $k \geq \mathrm{polylog}(n)$ | $\Theta(k)$ | $k^{-\Omega(1)}$ |
| This work | $O(1)$ | $k \geq \mathrm{polylog}(n)$ | $\Theta(k)$ | $n^{-\Omega(1)} + 2^{-k^{\Omega(1)}}$ |

Table 1: **Summary of Results on Extractors for Independent Sources.**

## 2 Overview of The Constructions and Techniques

Here we give a brief overview of our constructions and the techniques. To give a clear description of the ideas, we shall be informal and imprecise sometimes.

### 2.1 The high level idea

Similar as in [Rao06, BRSW06, Li13], our extractor is obtained by repeatedly condensing somewhere random sources. A somewhere random source (SR-source for short) is a random $N \times m$ matrix such that at least one row of the matrix is uniform. Given any $(n, k)$ weak source $X$, if we take a strong seeded extractor with seed length $d = O(\log n)$ and apply the extractor to $X$ using all possible choices of the seed, then the matrix obtained by concatenating all outputs of the seeded extractor is close to an SR-source with $N = 2^d = \text{poly}(n)$ rows.

The general idea of the condenser is to reduce the number of the rows in the SR-source in each step, while consuming a constant number of additional independent sources. When the number of rows is small enough (say $k^{O(1)}$), extraction becomes easy with another constant number of independent $(n, k)$ sources by using the extractor in [Rao06, BRSW06]. The condenser in [Rao06, BRSW06] reduces the number of the rows in the SR-source from $N$ to roughly $N/k^{0.9}$ in each step, while consuming a constant number of additional independent sources. Thus their extractor needs a total of $O(\log n/\log k)$ sources, and this performance is inherently limited by their techniques. Recently, Li [Li13] constructed a new condenser for SR-sources that breaks this limit. His condenser can reduce the number of the rows in the SR-source from $N$ to roughly $N^{3/4}$ in each step, while consuming one additional independent source. Thus, Li obtained an extractor that uses $O(\log(\frac{\log n}{\log k})) + O(1)$ sources. Our condenser follows the general paradigm suggested in [Li13]. Thus, we first describe the high-level ideas of the condenser in [Li13].

The condenser in [Li13] is based on a connection between the problem of condensing SR-sources and the problem of selecting a small committee in leader election. Namely, we can associate a player with each row in the SR-source. The players associated with uniform random rows can be viewed as honest players, and the players associated with the other rows can be viewed as faulty players, since their random bits can depend arbitrarily on the uniform random bits of the honest players. In this view the problem of reducing the number of rows in the SR-source is exactly the same as the problem of selecting a small committee of the players with enough honest players. Now suppose that a large fraction of the rows in the SR-source are uniform and independent, then this task can be done by using Feige's lightest bin protocol [Fei99]: pick $r$ bins and each player (each row in the SR-source) uses his random bits to select a bin randomly. The players (rows) that select the lightest bin form the selected committee (the new SR-source). Note that in the setting of leader election the faulty players can wait to see all the honest players' choices before making their choices, thus their random bits can indeed depend arbitrarily on the honest players' bits.

The idea of this simple and elegant protocol is as follows. Assume that the random bits of the honest players are uniform and independent, then by a Chernoff bound with high probability the honest players are roughly distributed evenly into each bin. When this happens, no matter how the faulty players make their choices, the lightest bin must contain at least roughly the same fraction of honest players as in the original set of players. In the leader election problem, assuming that there are at least a linear fraction of honest players whose random bits are uniform and independent, then in each step the number of players can be decreased from $N$ to roughly $\log N$ by using $N/\log N$ bins. In the next step the new set of players will use fresh random bits to perform the protocol again. In

the condenser problem, once we have selected a number of rows we take a strong seeded extractor and use the bits in the selected rows as seeds to extract random bits from another independent source. By the property of the strong extractor, as long as the length of the output is relatively short, with high probability the outputs of the "good" rows will remain uniform and independent, even conditioned on the previous SR-source. Thus, this new SR-source servers as fresh random bits for the players and we can run the lightest bin protocol again.

By applying a seeded extractor to a weak source and concatenating the outputs of all possible seeds, we can indeed obtain an SR-source such that a large fraction of the rows are (close to) uniform. However, they are not necessarily independent (in fact, it is impossible to make them all independent). Li [Li13] solved this problem by observing that for the lightest bin protocol to work, it suffices to have bounded independence instead of full independence. For example, if the uniform rows are pair-wise independence, then instead of using Chernoff's bound, one can use Chebysev's inequality to argue that again, with high probability the honest players are roughly distributed evenly into each bin, although we have to pick less bins in this case. By using a connection to non-malleable extractors and non-malleable condensers and the recent construction of non-malleable condensers for arbitrary min-entropy in [Li12a], Li indeed obtained an SR-source such that a large fraction of rows are (close to) pair-wise independent and uniform, from a constant number of independent sources. The pair-wise independence guarantees that the condenser can reduce the number of the rows in the SR-source from $N$ to roughly $N^{3/4}$ in each step.

Given the above discussion, one natural way to improve the condenser is to try to use higher orders of independence in the SR-source. Indeed, the more independent the uniform rows in the SR-source are, the faster the number of rows decreases in the condenser. In this paper, we indeed achieve this. Using a constant number of independent $(n, k)$ sources, we obtain an SR-source such that a large fraction of the rows are (close to) $h$-wise independent and uniform with $h = k^\alpha$ for some constant $0 < \alpha < 1$. Note that this is almost as good as possible, since the total entropy of the constant number of independent sources is $O(k)$, and the length of each row in the SR-source has to be at least $\log n$ in order to be used in a seeded extractor. Thus $h$ cannot be larger than say $k/\log n$. Once we have such an SR-source, we show that in each step the condenser can reduce the number of rows in the SR-source from $N$ to roughly $N^{4/\sqrt{h}}$, while consuming one more independent source. Thus, it only takes a constant number of independent sources for the number of rows to decrease to say $k^2$, even for entropy as small as $k = \text{polylog}(n)$.

## 2.2 Obtaining the $h$-wise independent SR-source

Thus, the remaining problem is to construct an SR-source such that a large fraction of the rows (at least a linear fraction) are (close to) $h$-wise independent and uniform. Note that we set $h = k^\alpha$ for some constant $0 < \alpha < 1$.

To do this, our first step is to control the error. We take a constant number of independent sources $X_1, \cdots, X_{\mathsf{C}}$ and a seeded extractor $\mathsf{Ext}$ with seed length $d = O(\log n)$, output length $m = 0.9k$ and error $1/\text{poly}(n)$, such as the extractor in [GUV09]. For every $i$, we apply $\mathsf{Ext}$ to $X_i$ using all possible seeds and concatenate the outputs to obtain a source $\bar{X}_i$ with $N = 2^d = \text{poly}(n)$ rows such that $1 - \epsilon$ fraction of the rows are $\epsilon$-close to uniform, for some $\epsilon = 1/\text{poly}(n)$. Now we compute the xor of all the $\bar{X}_i$'s and obtain a source $Y$. Note that there are at least $1 - \mathsf{C}\epsilon$ fraction of the rows in $Y$ such that the corresponding rows in all the $X_i$'s are $\epsilon$-close to uniform. Thus these rows in $Y$ are $\epsilon^{\mathsf{C}}$-close to uniform. Since $\epsilon = 1/\text{poly}(n)$ we can choose a constant $\mathsf{C}$ such that $\epsilon^{\mathsf{C}} < 1/N^2$. Thus we have that $Y$ is $N\epsilon^{\mathsf{C}} < 1/N$-close to an SR-source such that at least $1 - \mathsf{C}\epsilon$

fraction of the rows are *truly* uniform. We can now treat $Y$ as if it is this SR-source with a large fraction of truly uniform rows. This only adds error at most $1/N = 1/\text{poly}(n)$ to our extractor.

Next, we take one more independent source $X$ and we will obtain an SR-source $Z$ such that a large fraction of the rows are (close to) $h$-wise independent and uniform, from $X$ and $Y$. To do this, our starting point is the following alternating extraction protocol, which has been used a lot in recent constructions of non-malleable extractors and non-malleable condensers [DW09, Li12b, Li12a].

Specifically, assume that we have a weak source $X$ and a uniform string $Y^1$ independent of $X$. Take the first $\ell$ bits of $Y^1$ to be $S_1^1$, we compute the following random variables: $R_1^1 = \mathsf{Ext}(X, S_1^1), S_2^1 = \mathsf{Ext}(Y, R_1^1), R_2^1 = \mathsf{Ext}(X, S_2^1), \cdots, S_t^1 = \mathsf{Ext}(Y, R_{t-1}^1), R_t^1 = \mathsf{Ext}(X, S_t^1)$. Here $\mathsf{Ext}$ is a strong extractor that uses $\ell$ random bits to extract from a weak source and outputs $\ell$ bits as well, for some parameter $\ell$ that we will choose later. Thus in the above alternating extraction protocol, the size of each $S_i^1$ and $R_i^1$ is $\ell$. Now assume that we have $h-1$ random variables $Y^2, \cdots, Y^h$ (not necessarily uniform) which may be arbitrarily correlated with $Y$. If we run the alternating extraction protocol with $X$ and each $Y^i$, and for each $i$ obtain $t$ outputs $(R_1^i, \cdots, R_t^i)$, then one can show that as long as $X$ is independent of $(Y^1, \cdots, Y^h)$ and $\ell$ is small (roughly when the entropy of $X$ and the size of $Y$ is bigger than $ht\ell$), for any $1 \le j \le t$, we have that $R_j^1$ is close to uniform conditioned on $\{R_1^i, \cdots, R_{j-1}^i, i = 2, \cdots, h\}$.

Given this property, first consider the simple case where the SR-source $Y$ we obtained above has only $h$ rows. This is of course impossible since the number of rows in $Y$ is actually $N = \text{poly}(n) >> k > h$. Nevertheless, this simple hypothetical case would be illustrative to show our ideas. In this case, we can achieve (close to) true independence and uniform in the "good" rows of the source $Z$ obtained from $X$ and $Y$, as follows. For each row $Y^i$, run the alternating extraction protocol using $(X, Y^i)$ and output $h$ variables $R_1^i, \cdots, R_h^i$. Let $Z^i = R_i^i$ and $Z = Z^1 \circ \cdots \circ Z^h$. Now consider the rows which are uniform in $Y$. Their indexes can be ordered as $i_1 < i_2 < \cdots < i_t$ for some $t \le h$. Note that every row $Y^{i_j}$ is uniform, thus by the property of the alternating extraction, for any $1 \le j \le t$ we have that $Z^{i_j}$ is close to uniform conditioned on $(Z^{i_1}, \cdots, Z^{i_{j-1}})$. Therefore, $(Z^{i_1}, \cdots, Z^{i_t})$ is close to uniform.

It is now natural to generalize the above approach to the case where $Y$ has $N$ rows. However, the simple generalization does not work, since for each row it would require us to output $N$ variables $R_1^i, \cdots, R_N^i$ and there is not enough entropy in $X$ or $Y$ to make there $R$'s independent and uniform. On the other hand, we also do not need this since it implies that the good rows in $Z$ will be completely independent and uniform, while we only need $h$-wise independent. We solve this problem as follows. Note that the solution for the case where $Y$ has $h$ rows only consumes entropy roughly $h^2\ell$ (since we repeat $h$ times and each time produce $h$ random variables of size $\ell$). Thus, for the alternating extraction to work it suffices to take a slice of each $Y^i$ with size roughly $h^2\ell$. We will call this slice $Y^{i1}$. As above for each row $i$ we will use $Y^{i1}$ and $X$ to run an alternating extraction protocol to produce $h$ outputs $R_1^{i1}, \cdots, R_h^{i1}$. Now for each $i$, we divide the string that corresponds to $i$'s binary expression into equal blocks of size $\log h$ (padding 0's at the end if the last block does not have enough bits). Thus we get some $b = O(\log n / \log k)$ blocks. For each block we obtain an integer $\mathsf{Ind}_{ij}$ that corresponds to this block's binary expression. Now, for each row $i$ we will choose $R^{i1}_{\mathsf{Ind}_{i1}}$ as the output of the first step.

Now consider any $h$ uniform rows in $Y$. If their indexes in the first block happen to be all distinct then we are in good shape, since by the argument before we have that the corresponding $R^{i1}_{\mathsf{Ind}_{i1}}$'s will be independent and uniform. However, the indexes in the first block may not be all distinct. We have two observations here. First, for any row $v$ of these $h$ rows, $R^{v1}_{\mathsf{Ind}_{v1}}$ is close to

uniform conditioned on all the $R^{i1}_{\mathsf{Ind}_{i1}}$'s where the first block of $i$ is less than the first block of $v$. Second, we can fix all the $Y^{i1}$'s of the $h$ rows, and conditioned on this fixing, all $R^{i1}_{\mathsf{Ind}_{i1}}$'s are deterministic functions of $X$, and each $Y^i$ still has a lot of entropy (assuming that $k >> h^3\ell$). We will now take a strong seeded extractor $\mathsf{Ext}$ and for each row $i$, use $R^{i1}_{\mathsf{Ind}_{i1}}$ as a seed to extract from $Y^i$, where we obtain $Y^{i2} = \mathsf{Ext}(Y^i, R^{i1}_{\mathsf{Ind}_{i1}})$ with size roughly $h^2\ell$. The crucial observation here is that, we can now fix all $R^{i1}_{\mathsf{Ind}_{i1}}$'s, and conditioned on this fixing, all $Y^{i2}$'s are deterministic functions of $Y$, and $X$ still has a lot of entropy. Moreover, if for some row $v$ of these $h$ rows, $R^{v1}_{\mathsf{Ind}_{v1}}$ is close to uniform conditioned on all the $R^{i1}_{\mathsf{Ind}_{i1}}$'s with $i \in S_v$ for some subset of the rows $S_v$, then $Y^{v2}$ will be close to uniform conditioned on all the $Y^{i2}$'s with $i \in S_v$, as long as the size of each $Y^{i2}$ is not too large. Thus, at this point for each row $i$ we can use $Y^{i2}$ and $X$ to run another alternating extraction protocol to produce $h$ outputs $R^{i2}_1, \cdots, R^{i2}_h$, and choose $R^{i2}_{\mathsf{Ind}_{i2}}$ as the output of the second step.

The crucial observation here is that, since $Y^{v2}$ is close to uniform conditioned on all the $Y^{i2}$'s with $i \in S_v$ (where $S_v$ is the set of all $i$'s such that the first block of $i$ is less than the first block of $v$), we can first fix all the $Y^{i2}$'s with $i \in S_v$. Conditioned on this fixing, $Y^{v2}$ is still close to uniform. Moreover, conditioned on this fixing all the $R^{i2}_{\mathsf{Ind}_{i2}}$'s with $i \in S_v$ are deterministic functions of $X$. Thus we can further fix them. Conditioned on this fixing we have that $Y^{v2}$ is still close to uniform, $X$ still has a lot of entropy, and $X$ is still independent of the joint distribution of the unfixed $Y^{i2}$'s (since they are deterministic functions of $Y$). Therefore by the property of the alternating extraction, $R^{v2}_{\mathsf{Ind}_{v2}}$ will now be close to uniform even further conditioned on all the $R^{i2}_{\mathsf{Ind}_{i2}}$'s where the first block of $i$ is equal to the first block of $v$ but the second block of $i$ is less than the second block of $v$. Note that we have already fixed those $R^{i2}_{\mathsf{Ind}_{i2}}$'s where the first block of $i$ is less than the first block of $v$ before. Therefore we conclude that after the second step, for any row $v$ of these $h$ rows, $R^{v2}_{\mathsf{Ind}_{v2}}$ is close to uniform conditioned on all the $R^{i2}_{\mathsf{Ind}_{i2}}$'s where the integer corresponding to the first *two* blocks of $i$ is less than the integer corresponding to the first *two* block of $v$.

Now we can repeat the above procedure and use the third block to get $R^{i3}_{\mathsf{Ind}_{i3}}$, use the fourth block to get $R^{i4}_{\mathsf{Ind}_{i4}}$ and so on. When we reach the last block we will use $R^{ib}_{\mathsf{Ind}_{ib}}$ for the final block as our final output: $Z^i = R^{ib}_{\mathsf{Ind}_{ib}}$. Since the integer corresponding to all the $b$ blocks of $i$ is just $i$, we now have that for any row $v$ of these $h$ rows, $R^{vb}_{\mathsf{Ind}_{vb}}$ is close to uniform conditioned on all the $R^{ib}_{\mathsf{Ind}_{ib}}$'s where $i < v$. Therefore, the joint distribution of the $Z^i$'s of the $h$ rows is close to uniform, as we desire. Note that this requires that $k > bh^3\ell$ roughly. We also need to control the error in this process. For this we will choose $\ell = k^\beta$ for some constant $0 < \beta < 1$ with $\beta > \alpha$. By using a seeded extractor with seed length $O(\log n + \log(1/\epsilon))$ such as that in [GUV09], the total error of this process can be bounded as $\epsilon = O(bh^2 2^{-\Omega(\ell)}) = 2^{-\Omega(\ell)}$. By choosing $\alpha, \beta$ appropriately we can ensure that $\epsilon << N^{-h}$, which is good enough for the lightest bin protocol. Note that $b = O(\log n / \log k)$, $h = k^\alpha$ and $\ell > \log n$, this implies that $k > bh^3\ell > \log^2 n$. Thus our extractor can only work for weak sources with min-entropy $k \geq \log^{2+\eta} n$ for any constant $\eta > 0$.

### 2.3 Comparison to the construction in [Li13]

As can be seen from the above discussion, the most technical part in our construction is to obtain an SR-source such that a large fraction of the rows are close to $h$-wise independent and uniform. This is also true for the construction in [Li13]. However, the SR-source in that construction only achieves pair-wise independence. Here, we outline two key differences between our construction and the construction in [Li13], which enable us to overcome the difficulties in [Li13] and finally

achieve $h$-wise independence and uniform with $h = k^\alpha$ for some constant $0 < \alpha < 1$, using only a constant number of sources.

First, we deal with the error in a different way. The construction in [Li13] first uses ideas related to non-malleable condensers to get an SR-source such that for a large fraction of the rows, each pair is $\epsilon$-close to being independent and uniform. Limited by the use of a seeded extractor with seed length $O(\log n)$ (when we try all the possible seeds), the error $\epsilon$ is $1/\text{poly}(n) > 1/N$. However, for the lightest bin protocol to work the error needs to be roughly $N^{-h}$ in the case of $h$-wise independence. The construction in [Li13] deals with this by preparing independent copies of the pair-wise independent SR-sources and taking the xor of them. If we take $t$ copies then the error reduces to $\epsilon^t$. This is enough for the pair-wise independence case since it only requires a constant number of copies. However, if we want to generalize to $h$-wise independence then this approach would require $O(h)$ sources, which we cannot afford for a super constant $h$.

In this paper, we instead control the error before constructing the SR-source with $h$-wise independent property. We first take a constant number $\mathsf{C}$ of independent sources and from each of them obtain an SR-source with $N$ rows (without the $h$-wise independent property) such that a large fraction of the rows are $\epsilon$-close to uniform for some $\epsilon = 1/\text{poly}(n) > 1/N$. We then take the xor of these sources so that the error in the "good" rows reduces to $\epsilon^{\mathsf{C}} < 1/N^2$. Now we can show that the new SR-source $Y$ is (globally) $N\epsilon^{\mathsf{C}} = 1/\text{poly}(n)$-close to an SR-source $Y'$ where a large fraction of the rows are truly uniform. Therefore, we can now treat $Y$ as $Y'$. This gives us uniform bits in the good rows, so that we can use $\ell = k^\beta$ bits to achieve error $2^{-\Omega(\ell)}$. For a carefully chosen $\beta > \alpha$ this error is good enough for the lightest bin protocol.

Second, the SR-source constructed in [Li13] is actually stronger than what we need. Indeed, by using ideas related to non-malleable condenser, not only do we get pair-wise independence in the good rows, but also the output of any good row is close to uniform conditioned on the output of any other single row. This stronger property does not seem easy to generalize to $h$-wise independence for large $h$. In this paper, instead, we only achieve $h$-wise independence in the good rows, which is all we need for the lightest bin protocol to work. This allows us to use the alternating extraction more directly. For example, in the hypothetical case where $Y$ has only $h$ rows, we use each row to do the alternating extraction and produce $h$ $R_i$'s. Now for each row $i$ we pick $R_i^i$ (i.e., the diagonal elements). Since the alternating extraction guarantees that as long as the row $Y^i$ is uniform, $R_i^i$ is close to uniform conditioned on all the $R_j^j$'s with $j < i$, this ensures that the joint distribution of all outputs of the good rows is close to uniform. On the other hand, it may not be true that $R_i^i$ is close to uniform conditioned on the output of a bad row $R_j^j$ if $j > i$.

**Organization.** The rest of the paper is organized as follows. We give some preliminaries in Section 3. In Section 4 we define alternating extraction, an important ingredient in our construction. We present our main construction of extractors in Section 5. Finally we conclude with some open problems in Section 6.

## 3 Preliminaries

We often use capital letters for random variables and corresponding small letters for their instantiations. Let $|S|$ denote the cardinality of the set $S$. For $\ell$ a positive integer, $U_\ell$ denotes the uniform distribution on $\{0,1\}^\ell$. When used as a component in a vector, each $U_\ell$ is assumed independent of the other components. All logarithms are to the base 2.

### 3.1 Probability distributions

**Definition 3.1** (statistical distance)**.** Let $W$ and $Z$ be two distributions on a set $S$. Their *statistical distance* (variation distance) is

$$\Delta(W,Z) \stackrel{def}{=} \max_{T\subseteq S}(|W(T) - Z(T)|) = \frac{1}{2}\sum_{s\in S}|W(s) - Z(s)|.$$

We say $W$ is $\varepsilon$-close to $Z$, denoted $W \approx_\varepsilon Z$, if $\Delta(W,Z) \leq \varepsilon$. For a distribution $D$ on a set $S$ and a function $h : S \to T$, let $h(D)$ denote the distribution on $T$ induced by choosing $x$ according to $D$ and outputting $h(x)$.

### 3.2 Somewhere Random Sources and Extractors

**Definition 3.2** (Somewhere Random sources)**.** A source $X = (X_1, \cdots, X_t)$ is $(t \times r)$ *somewhere-random* (SR-source for short) if each $X_i$ takes values in $\{0,1\}^r$ and there is an $i$ such that $X_i$ is uniformly distributed.

**Definition 3.3.** (Seeded Extractor) A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a *strong* $(k,\varepsilon)$-*extractor* if for every source $X$ with min-entropy $k$ and independent $Y$ which is uniform on $\{0,1\}^d$,

$$(\mathsf{Ext}(X,Y), Y) \approx_\varepsilon (U_m, Y).$$

### 3.3 Average conditional min-entropy

**Definition 3.4.** The *average conditional min-entropy* is defined as

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

**Lemma 3.5** ([DORS08])**.** *For any $s > 0$, $\Pr_{w\leftarrow W}[H_\infty(X|W = w) \geq \widetilde{H}_\infty(X|W) - s] \geq 1 - 2^{-s}$.*

**Lemma 3.6** ([DORS08])**.** *If a random variable $B$ has at most $2^\ell$ possible values, then $\widetilde{H}_\infty(A|B) \geq H_\infty(A) - \ell$.*

### 3.4 Prerequisites from previous work

Sometimes it is convenient to talk about average case seeded extractors, where the source $X$ has average conditional min-entropy $\widetilde{H}_\infty(X|Z) \geq k$ and the output of the extractor should be uniform given $Z$ as well. The following lemma is proved in [DORS08].

**Lemma 3.7.** *[DORS08] For any $\delta > 0$, if $\mathsf{Ext}$ is a $(k,\epsilon)$ extractor then it is also a $(k+\log(1/\delta), \epsilon+\delta)$ average case extractor.*

For a strong seeded extractor with optimal parameters, we use the following extractor constructed in [GUV09].

**Theorem 3.8** ([GUV09])**.** *For every constant $\alpha > 0$, and all positive integers $n, k$ and any $\epsilon > 0$, there is an explicit construction of a strong $(k,\epsilon)$-extractor $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log n + \log(1/\epsilon))$ and $m \geq (1-\alpha)k$.*

**Theorem 3.9** ([Rao06, BRSW06]). *There exist constants $c > 0$ and $c'$ such that for every $n, k$ with $k = k(n) > \log^2 n$ and $\ell \leq \text{poly}(n)$ there exists a polynomial time computable function $\mathsf{SRExt} : \{0,1\}^{\ell k} \times \{0,1\}^{un} \to \{0,1\}^m$ with $m = \Omega(k)$ and $u \leq c' \frac{\log \ell}{\log k}$ s.t. if $X^1, X^2 \cdots, X^u$ are independent $(n,k)$ sources and $Y$ is an independent $\ell \times k$ SR-source then*

$$|\mathsf{SRExt}(Y, X^1, X^2 \cdots, X^u) - U_m| < 2^{-k^c}.$$

The following standard lemma about conditional min-entropy is implicit in [NZ96] and explicit in [MW97].

**Lemma 3.10** ([MW97]). *Let $X$ and $Y$ be random variables and let $\mathcal{Y}$ denote the range of $Y$. Then for all $\epsilon > 0$, one has*

$$\Pr_Y \left[ H_\infty(X|Y = y) \geq H_\infty(X) - \log |\mathcal{Y}| - \log \left( \frac{1}{\epsilon} \right) \right] \geq 1 - \epsilon.$$

We also need the following lemma.

**Lemma 3.11.** *[Li12a] Let $(X, Y)$ be a joint distribution such that $X$ has range $\mathcal{X}$ and $Y$ has range $\mathcal{Y}$. Assume that there is another random variable $X'$ with the same range as $X$ such that $|X - X'| = \epsilon$. Then there exists a joint distribution $(X', Y)$ such that $|(X, Y) - (X', Y)| = \epsilon$.*

**Lemma 3.12.** *[BIW04] Assume that $Y_1, Y_2, \cdots, Y_t$ are independent random variables over $\{0,1\}^n$ such that for any $i, 1 \leq i \leq t$, we have $|Y_i - U_n| \leq \epsilon$. Let $Z = \oplus_{i=1}^t Y_i$. Then $|Z - U_n| \leq \epsilon^t$.*

# 4 Alternating Extraction

An important ingredient in our construction is the following alternating extraction protocol.

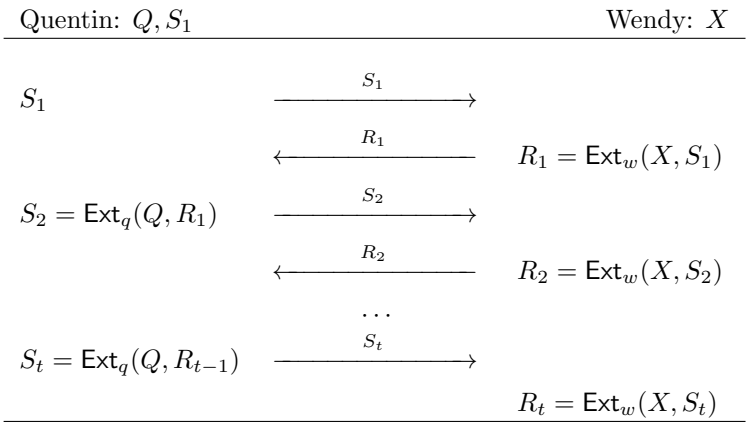


Figure 1: Alternating Extraction.

**Alternating Extraction.** Assume that we have two parties, Quentin and Wendy. Quentin has a source $Q$, Wendy has a source $X$. Also assume that Quentin has a uniform random seed $S_1$ (which may be correlated with $Q$). Suppose that $(Q, S_1)$ is kept secret from Wendy and $X$ is kept secret from Quentin. Let $\mathsf{Ext}_q$, $\mathsf{Ext}_w$ be strong seeded extractors with optimal parameters, such as

that in Theorem 3.8. Let $\ell$ be an integer parameter for the protocol. For some integer parameter $t > 0$, the *alternating extraction protocol* is an interactive process between Quentin and Wendy that runs in $t$ steps.

In the first step, Quentin sends $S_1$ to Wendy, Wendy computes $R_1 = \mathsf{Ext}_w(X, S_1)$. She sends $R_1$ to Quentin and Quentin computes $S_2 = \mathsf{Ext}_q(Q, R_1)$. In this step $R_1, S_2$ each outputs $\ell$ bits. In each subsequent step $i$, Quentin sends $S_i$ to Wendy, Wendy computes $R_i = \mathsf{Ext}_w(X, S_i)$. She replies $R_i$ to Quentin and Quentin computes $S_{i+1} = \mathsf{Ext}_q(Q, R_i)$. In step $i$, $R_i, S_{i+1}$ each outputs $\ell$ bits. Therefore, this process produces the following sequence:

$$S_1, R_1 = \mathsf{Ext}_w(X, S_1), S_2 = \mathsf{Ext}_q(Q, R_1), \cdots, S_t = \mathsf{Ext}_q(Q, R_{t-1}), R_t = \mathsf{Ext}_w(X, S_t).$$

**Look-Ahead Extractor.** Now we can define our look-ahead extractor. Let $Y = (Q, S_1)$ be a seed, the look-ahead extractor is defined as

$$\mathsf{laExt}(X, Y) = \mathsf{laExt}(X, (Q, S_1)) \stackrel{def}{=} R_1, \cdots, R_t.$$

We first prove the following lemma.

**Lemma 4.1.** *Let $Y = (Q, S_1)$ where $Q$ is an $(n_q, k_q)$ source and $S_1$ is the uniform distribution over $\ell$ bits. Let $Y_2 = (Q_2, S_{21}), \cdots, Y_h = (Q_h, S_{h1})$ be another $h-1$ random variables with the same range of $Y$ that are arbitrarily correlated to $Y$. Assume that $X$ is an $(n, k)$ source independent of $(Y, Y_2, \cdots, Y_h)$, such that $k > ht\ell + 10\ell + 2\log(1/\epsilon)$ and $k_q > ht\ell + 10\ell + 2\log(1/\epsilon)$. Assume that $\mathsf{Ext}_q$ and $\mathsf{Ext}_w$ are strong seeded extractors that use $\ell$ bits to extract from $(n_q, 10\ell)$ sources and $(n, 10\ell)$ sources respectively, with error $\epsilon$ and $\ell = O(\log(max\{n_q, n\}) + \log(1/\epsilon))$. Let $(R_1, \cdots, R_t) = \mathsf{laExt}(X, Y)$ and $(R_{i1}, \cdots, R_{it}) = \mathsf{laExt}(X, Y_i)$ for $i = 2, \cdots, h$. Then for any $0 \le j \le t-1$, we have*

$$(Y, Y_2, \cdots, Y_h, \{R_{i1}, \cdots, R_{ij}, i = 2, \cdots, h\}, R_{j+1}) \approx_{\epsilon_1} (Y, Y_2, \cdots, Y_h, \{R_{i1}, \cdots, R_{ij}, i = 2, \cdots, h\}, U_\ell),$$

*where $\epsilon_1 = O(t\epsilon)$.*

*Proof.* For any $i = 2, \cdots, h$, let $\{S_{ij}, j = 1, \cdots, t\}$ denote the random variables corresponding to $\{S_j\}$ that are produced in $\mathsf{laExt}(X, Y_i)$. For any $j, 1 \le j \le t$, let $\overline{S_j} = (S_1, \cdots, S_j)$ and $\overline{S_{ij}} = (S_{i1}, \cdots, S_{ij})$ for $i = 2, \cdots, h$. Let $\overline{R_j} = (R_1, \cdots, R_j)$ and $\overline{R_{ij}} = (R_{i1}, \cdots, R_{ij})$ for $i = 2, \cdots, h$. We prove the following stronger claim.

**Claim 4.2.** *For any $j$, we have that*

$$\begin{aligned}&(R_j, \overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_{j-1}}, \{\overline{R_{i(j-1)}}, i = 2, \cdots, h\}, Y, Y_2, \cdots, Y_h)\\ \approx_{(4j-2)\epsilon}&(U_\ell, \overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_{j-1}}, \{\overline{R_{i(j-1)}}, i = 2, \cdots, h\}, Y, Y_2, \cdots, Y_h).\end{aligned}$$

*and*

$$\begin{aligned}&(S_{j+1}, \overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\})\\ \approx_{(4j)\epsilon}&(U_\ell, \overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}).\end{aligned}$$

*Moreover, conditioned on* $(\overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_{j-1}}, \{\overline{R_{i(j-1)}}, i = 2, \cdots, h\})$, $(R_j, \{R_{ij}, i = 2, \cdots, h\})$ *are deterministic functions of* $X$ *and the average conditional min-entropy of* $Q$ *is at least* $k_q - hj\ell$*; conditioned on* $(\overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\})$, $(Q, Q_2, \cdots, Q_h, S_{j+1}, \{S_{i(j+1)}, i = 2, \cdots, h\})$ *is independent of* $X$ *and the average conditional min-entropy of* $X$ *is at least* $k - hj\ell$*.*

We prove the claim by induction on $j$. When $j = 0$, the statement is trivially true. Now we assume that the statements hold for some $j$ and we prove them for $j + 1$.

We first fix $(\overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\})$. Note that now $(Q, Q_2, \cdots, Q_h, S_{j+1}, \{S_{i(j+1)}, i = 2, \cdots, h\})$ is independent of $X$. Moreover conditioned on this fixing $S_{j+1}$ is $(4j)\epsilon$-close to uniform on average. Note that the average conditional min-entropy of $X$ is at least $k - hj\ell \geq k - ht\ell > 10\ell + 2\log(1/\epsilon)$. By Theorem 3.8 and Lemma 3.7 we have that $\mathsf{Ext}_w$ is a $(10\ell + \log(1/\epsilon), 2\epsilon)$ average case strong extractor. Thus

$$\begin{aligned}&(R_{j+1}, \overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}, S_{j+1})\\ \approx_{(4j+2)\epsilon}&(U_\ell, \overline{S_j}, \{\overline{S_{ij}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}, S_{j+1}).\end{aligned}$$

Since $(Q, Q_2, \cdots, Q_h, S_{j+1}, \{S_{i(j+1)}, i = 2, \cdots, h\})$ is independent of $X$, and $R_{j+1}$ is a deterministic function of $X$ conditioned on $S_{j+1}$, we also have

$$\begin{aligned}&(R_{j+1}, \overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}, Y, Y_2, \cdots, Y_h)\\ \approx_{(4j+2)\epsilon}&(U_\ell, \overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}, Y, Y_2, \cdots, Y_h).\end{aligned}$$

Moreover, conditioned on $(\overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\})$, $(R_{j+1}, \{R_{i(j+1)}, i = 2, \cdots, h\})$ are deterministic functions of $X$, and the average conditional min-entropy of $Q$ is at least $k_q - hj\ell - h\ell = k_q - h(j + 1)\ell$.

Next, since conditioned on $(\overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\})$, $(R_{j+1}, \{R_{i(j+1)}, i = 2, \cdots, h\})$ are deterministic functions of $X$, they are independent of $(Q, Q_2, \cdots, Q_h)$. Moreover conditioned on this fixing $R_{j+1}$ is $(4j + 2)\epsilon$-close to uniform on average. Note that the average conditional min-entropy of $Q$ is at least $k_q - h(j+1)\ell \geq k_q - ht\ell > 10\ell + 2\log(1/\epsilon)$. By Theorem 3.8 and Lemma 3.7 we have that $\mathsf{Ext}_q$ is a $(10\ell + \log(1/\epsilon), 2\epsilon)$ average case strong extractor. Thus

$$\begin{aligned}&(S_{j+2}, \overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}, R_{j+1})\\ \approx_{(4j+4)\epsilon}&(U_\ell, \overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_j}, \{\overline{R_{ij}}, i = 2, \cdots, h\}, R_{j+1}).\end{aligned}$$

Since $(R_{j+1}, \{R_{i(j+1)}, i = 2, \cdots, h\})$ are independent of $(Q, Q_2, \cdots, Q_h)$, and $S_{j+2}$ is a deterministic function of $Q$ conditioned on $R_{j+1}$, we also have

$$\begin{aligned}&(S_{j+2}, \overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_{j+1}}, \{\overline{R_{i(j+1)}}, i = 2, \cdots, h\})\\ \approx_{(4j+4)\epsilon}&(U_\ell, \overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_{j+1}}, \{\overline{R_{i(j+1)}}, i = 2, \cdots, h\}).\end{aligned}$$

Moreover, conditioned on $(\overline{S_{j+1}}, \{\overline{S_{i(j+1)}}, i = 2, \cdots, h\}, \overline{R_{j+1}}, \{\overline{R_{i(j+1)}}, i = 2, \cdots, h\})$, $(S_{j+2}, \{S_{i(j+2)}, i = 2, \cdots, h\})$ are deterministic functions of $(Q, Q_2, \cdots, Q_h)$. Thus $(Q, Q_2, \cdots, Q_h, S_{j+2}, \{S_{i(j+2)}, i = 2, \cdots, h\})$ is independent of $X$. The average conditional min-entropy of $X$ is at least $k - hj\ell - h\ell = k - h(j + 1)\ell$.

Note that $j \leq t$, thus the lemma is proved. □

## 5 The Extractor

In this section we give our main construction. The first step is to obtain a source that is close to an SR-source such that a large fraction of the rows are uniform. We have the following algorithm.

**Algorithm 5.1** ($\mathsf{SR}(X_1, \cdots, X_{\mathsf{C}})$)**.**

**Input:** $\mathsf{C}$–an integer constant. $X_1, \cdots, X_{\mathsf{C}}$ — independent $(n, k)$-sources with $k \geq \text{polylog}(n)$.
**Output:** $Y$ — a source that is close to an SR-source.

**Sub-Routines and Parameters:**
Let $\mathsf{Ext}$ be the strong extractor with optimal parameters from Theorem 3.8, with error $\epsilon' = 1/\text{poly}(n)$ and seed length $d = O(\log n)$, set up to output $0.9k$ bits.

1. For every $i = 1, \cdots, \mathsf{C}$ do the following. For every $r \in \{0,1\}^d$ compute $X_i^j = \mathsf{Ext}(X_i, r)$, where $j-1$ is the integer whose binary expression is $r$. Let $\overline{X}_i = X_i^1 \circ \cdots \circ X_i^N$ where $N = 2^d = \text{poly}(n)$.
2. Compute $Y = \bigoplus_{i=1}^{\mathsf{C}} \overline{X}_i$.

We have the following lemma.

**Lemma 5.2.** *There exists a constant integer $\mathsf{C} > 1$ and $\epsilon = 1/\text{poly}(n)$ such that $Y$ is $\epsilon$-close to another source $Y' = Y'^1 \circ \cdots \circ Y'^N$ where each $Y'^i$ has $0.9k$ bits, and the following holds. There exists a subset $S \subset [N]$ with $|S| > 0.9N$ such that for any $i \in S$, $Y'^i$ is uniform.*

*Proof.* By Theorem 3.8, for every $X_i$ there exists a subset $S_i \subset [N]$ with $|S_i| \geq (1-\epsilon_1)N$ such that for any $j \in S_i$, we have $X_i^j = \mathsf{Ext}(X_i, r)$ is $\epsilon_1$-close to uniform, where $\epsilon_1 = \sqrt{\epsilon'} = 1/\text{poly}(n)$. Since $N = 2^d = \text{poly}(n)$, there exists a constant integer $\mathsf{C} > 1$ such that $\epsilon_1^{\mathsf{C}} < 1/N^2$. Now take $\mathsf{C}$ independent sources $X_1, \cdots, X_{\mathsf{C}}$ and let $Y = Y^1 \circ \cdots \circ Y^N$ be obtained as above. Thus for any $j \in [N]$ we have $Y = \bigoplus_{i=1}^{\mathsf{C}} X_i^j$.

Let $S = \cap S_i$. Thus $|S| \geq 1 - \mathsf{C}\epsilon_1 > 0.9N$. Note that for any $j \in S$, we have that $\forall i, X_i^j$ is $\epsilon_1$-close to uniform. Thus by Lemma 3.12, for any $j \in S$ we have that $Y^j$ is $\epsilon_2 = \epsilon_1^{\mathsf{C}}$-close to uniform. Now by Lemma 3.11, we can change $\{Y^j, j \in S\}$ one by one to the uniform distribution. In each step we only change one $Y^j$ while keeping the joint distribution of all the other rows fixed. In the end the source $Y$ is changed into another source $Y' = Y'^1 \circ \cdots \circ Y'^N$ such that for any $i \in S$, $Y'^i$ is uniform. Since in each step the statistical distance is at most $\epsilon_2$, the statistical distance between $Y$ and $Y'$ is at most $N\epsilon_2 < 1/N = 1/\text{poly}(n)$ by the triangle inequality. □

The next step is to obtain an SR-source such that a large fraction of the rows are roughly $h$-wise independent with $h = k^\alpha$ for some constant $0 < \alpha < 1$. For this we have the following algorithm.

**Algorithm 5.3** ($\mathsf{SSR}(X, Y)$)**.**

**Input:** $X$— an $(n, k)$-source with $k \geq \text{polylog}(n)$. $Y = Y^1 \circ \cdots \circ Y^N$—an SR-source with $N = \text{poly}(n)$ rows and each row has $0.9k$ bits, independent of $X$.
**Output:** $Z$ — a source that is close to an SR-source.

**Sub-Routines and Parameters:**
Let $0 < \alpha < \beta < 1$ be two constants to be chosen later. Let $\ell = k^\beta$. Pick an integer $h$ such that $k^\alpha \leq h < 2k^\alpha$ and $h = 2^l$ for some integer $l > 0$. Let $\mathsf{Ext}_q, \mathsf{Ext}_w$ be strong extractors with optimal parameters from Theorem 3.8, set up to extract from $((h^2+12)\ell, 10\ell)$ sources and $(n, 10\ell)$ sources respectively, with seed length $\ell$, error $\epsilon = 2^{-\Omega(\ell)}$ and output length $\ell$. These will be used in $\mathsf{laExt}$. Let $\mathsf{Ext}$ be a strong extractor with optimal parameters from Theorem 3.8, set up to extract from $(0.9k, 2(h^2+12)\ell)$ sources, with seed length $\ell$, error $\epsilon = 2^{-\Omega(\ell)}$ and output length $(h^2+12)\ell$.

1. For every $i = 1, \cdots, N$, use $X$ and $Y^i$ to compute $Z^i$ as follows.
   (a) Compute the binary expression of $i-1$, which consists of $d = \log N = O(\log n)$ bits. Divide these bits sequentially from left to right into $b = \lceil \frac{d}{l} \rceil$ blocks of size $l$ (the last block may have less than $l$ bits, then we add 0s at the end to make it $l$ bits). Now from left to right, for each block $j = 1, \cdots, b$, we obtain an integer $\mathsf{Ind}_{ij} \leq 2^l$ such that the binary expression of $\mathsf{Ind}_{ij} - 1$ is the same as the bits in block $j$.
   (b) Let $Y^{i1}$ be the first $(h+12)\ell$ bits of $Y^i$. Set $j = 1$. While $j < b$ do the following.
      i. Compute $(R_1^{ij}, \cdots, R_h^{ij}) = \mathsf{laExt}(X, Y^{ij})$, where $Q = Y^{ij}$ and $S_1$ is the first $\ell$ bits of $Y^{ij}$.
      ii. Compute $Y^{i(j+1)} = \mathsf{Ext}(Y^i, R^{ij}_{\mathsf{Ind}_{ij}})$.
      iii. Set $j = j + 1$.
   (c) Finally, compute $(R_1^{ib}, \cdots, R_h^{ib}) = \mathsf{laExt}(X, Y^{ib})$ and set $Z^i = R^{ib}_{\mathsf{Ind}_{ib}}$.
2. Let $Z = Z^1 \circ \cdots \circ Z^N$.

We now introduce some notation. For any $i \in [N]$ and $j \in [b]$, we let $Y^{i(\leq j)}$ denote $(Y^{i1}, \cdots, Y^{ij})$, let $R^{i(\leq j)}_{\mathsf{Ind}_{i(\leq j)}}$ denote $(R^{i1}_{\mathsf{Ind}_{i1}}, \cdots, R^{ij}_{\mathsf{Ind}_{ij}})$ and let $f^j(i)$ denote the integer whose binary expression is the concatenation of the binary expression of $i-1$ from block 1 to block $j$. We have the following lemma.

**Lemma 5.4.** *Assume that $k \geq 2(bh+2)(h^2+12)\ell$. Fix any $v \in [N]$ such that $Y^v$ is uniform. Let $S \subset [N]$ be any subset with $|S| = h$ and $v \in S$. For any $j \in [b]$, define $S_v^j = \{i \in S : f^j(i) < f^j(v)\}$. Then for any $j \in [b]$, we have that*

$$(R^{vj}_{\mathsf{Ind}_{vj}}, \{Y^{i(\leq j)}, i \in S\}, \{R^{ij}_{\mathsf{Ind}_{ij}}, i \in S_v^j\}, \{R^{i(\leq j-1)}_{\mathsf{Ind}_{i(\leq j-1)}}, i \in S\})$$
$$\approx_{O(jh\epsilon)} (U_\ell, \{Y^{i(\leq j)}, i \in S\}, \{R^{ij}_{\mathsf{Ind}_{ij}}, i \in S_v^j\}, \{R^{i(\leq j-1)}_{\mathsf{Ind}_{i(\leq j-1)}}, i \in S\}).$$

*Moreover, conditioned on the fixing of* $(\{Y^{i(\le j)}, i \in S\}, \{R^{i(\le j-1)}_{\mathsf{Ind}_{i(\le j-1)}}, i \in S\})$*, we have that*

1. *$X$ and $Y$ are still independent.*
2. *$(R^{ij}_{\mathsf{Ind}_{ij}}, i \in S)$ are all deterministic functions of $X$.*
3. *The average conditional min-entropy of $X$ is at least $k-(j-1)h\ell$ and the average conditional min-entropy of $Y^v$ is at least $0.9k - jh(h^2+12)\ell$.*

*Proof.* We prove the lemma by induction on $j$. When $j = 1$, note that $f^j(i) = f^1(i) = \mathsf{Ind}_{i1} - 1$ and $Y^{i(\le 1)} = Y^{i1}$. Note that $S^1_v$ contains all the $i \in S$ such that $\mathsf{Ind}_{i1} < \mathsf{Ind}_{v1}$. Thus the first part of the lemma follows directly from Lemma 4.1 (note that $t = h$ in Lemma 4.1). Moreover, since $\{Y^{i1}, i \in S\}$ are deterministic functions of $Y$, conditioned on the fixing of them we have $X$ and $Y$ are still independent, the average conditional min-entropy of $Y^v$ is at least $0.9k - h(h^2+12)\ell$ and the min-entropy of $X$ is $k$. Note also that after this fixing, $(R^{i1}_{\mathsf{Ind}_{i1}}, i \in S)$ are all deterministic functions of $X$. So the lemma holds for $j = 1$.

Now assume that the lemma holds for some $j \le b-1$, we show that it also holds for $j+1$. We first fix $(\{Y^{i(\le j)}, i \in S\}, \{R^{i(\le j-1)}_{\mathsf{Ind}_{i(\le j-1)}}, i \in S\})$. By the induction hypothesis we have that conditioned on this fixing, $X$ and $Y$ are still independent, the average conditional min-entropy of $X$ is at least $k-(j-1)h\ell$ and the average conditional min-entropy of $Y^v$ is at least $0.9k - jh(h^2+12)\ell$, and that $(R^{ij}_{\mathsf{Ind}_{ij}}, i \in S)$ are all deterministic functions of $X$.

Now consider the set $S^{j+1}_v = \{i \in S : f^{j+1}(i) < f^{j+1}(v)\}$. It is easy to see that $S^j_v \subseteq S^{j+1}_v$. Let $\overline{S}^j_v = S^{j+1}_v \setminus S^j_v$. In other words, $\overline{S}^j_v$ contains all the $i \in [S]$ such that the first $j$ blocks of $i-1$'s binary expression are the same as the first $j$ blocks of $v-1$'s binary expression, but the $j+1$'th block of $i-1$'s binary expression is smaller than the $j+1$'th block of $v-1$'s binary expression. Note that it is possible that $\overline{S}^j_v = \phi$. However, if $\overline{S}^j_v \neq \phi$ then for any $i \in \overline{S}^j_v$ we have that $\mathsf{Ind}_{i(j+1)} < \mathsf{Ind}_{v(j+1)}$.

We now further fix $(R^{ij}_{\mathsf{Ind}_{ij}}, i \in S^j_v)$. By the induction hypothesis, conditioned on this fixing $R^{vj}_{\mathsf{Ind}_{vj}}$ is still $O(jh\epsilon)$-close to uniform on average. Moreover, $(R^{ij}_{\mathsf{Ind}_{ij}}, i \in S \setminus S^j_v)$ are still functions of $X$ and are thus independent of $Y$. Now conditioned on this fixing, we have that $(Y^{i(j+1)} = \mathsf{Ext}(Y^i, R^{ij}_{\mathsf{Ind}_{ij}}), i \in S^j_v)$ are deterministic functions of $Y$. Thus we can further fix $(Y^{i(j+1)}, i \in S^j_v)$ and conditioned on this fixing, $X$ and $Y$ are still independent. Moreover, the average conditional min-entropy of $Y^v$ is at least $0.9k - jh(h^2+12)\ell - |S^j_v|(h^2+12)\ell \ge 0.9k - (j+1)h(h^2+12)\ell \ge 0.9k - bh(h^2+12)\ell \ge 2(h^2+12)\ell$. Thus by Theorem 3.8 and Lemma 3.7 we have that

$$\begin{aligned}&(Y^{v(j+1)}, \{Y^{i(j+1)}, i \in S^j_v\}, R^{vj}_{\mathsf{Ind}_{vj}}, \{R^{ij}_{\mathsf{Ind}_{ij}}, i \in S^j_v\})\\ \approx_{O(jh\epsilon)+2\epsilon}&(U_{(h^2+12)\ell}, \{Y^{i(j+1)}, i \in S^j_v\}, R^{vj}_{\mathsf{Ind}_{vj}}, \{R^{ij}_{\mathsf{Ind}_{ij}}, i \in S^j_v\}).\end{aligned}$$

Since conditioned on $R^{vj}_{\mathsf{Ind}_{vj}}$, $Y^{v(j+1)}$ is a deterministic function of $Y$, and $(R^{ij}_{\mathsf{Ind}_{ij}}, i \in S \setminus S^j_v)$ are functions of $X$, we also have that

$$\begin{aligned}&(Y^{v(j+1)}, \{Y^{i(j+1)}, i \in S^j_v\}, \{R^{ij}_{\mathsf{Ind}_{ij}}, i \in S\})\\ \approx_{O(jh\epsilon)+2\epsilon}&(U_{(h^2+12)\ell}, \{Y^{i(j+1)}, i \in S^j_v\}, \{R^{ij}_{\mathsf{Ind}_{ij}}, i \in S\}).\end{aligned}$$

Now we fix all $(R^{ij}_{\mathsf{Ind}_{ij}}, i \in S)$. Since these are all deterministic functions of $X$, conditioned on this fixing $X$ and $Y$ are still independent. Moreover, the average conditional min-entropy of $X$ is at least $k-(j-1)h\ell-h\ell = k-jh\ell$. Note that conditioned on this fixing, all $(Y^{i(j+1)}, i \in S)$ are deterministic functions of $Y$, and are thus independent of $X$. Note that we have fixed $(Y^{i(j+1)}, i \in S^j_v)$ before. The equation above implies that conditioned on all these fixings, $Y^{v(j+1)}$ is still $O(jh\epsilon)+2\epsilon$-close to uniform on average and independent of $X$. Next, note that after these fixings, $(R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^j_v)$ are all deterministic functions of $X$. Thus we can now further fix $(R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^j_v)$, and conditioned on this fixing, $X$ and $Y$ are still independent, thus $X$ is also independent of $(Y^{i(j+1)}, i \in S \setminus S^j_v)$. Note that the average conditional min-entropy of $X$ is at least $k-jh\ell-|S^j_v|\ell \geq k-(j+1)h\ell \geq k-bh\ell \geq (h^2+12)\ell$. Note that if $\overline{S}^j_v \neq \phi$ then for any $i \in \overline{S}^j_v$ we have that $\mathsf{Ind}_{i(j+1)} < \mathsf{Ind}_{v(j+1)}$. Therefore by Lemma 4.1 we have that

$$\begin{aligned}&(R^{v(j+1)}_{\mathsf{Ind}_{v(j+1)}}, \{R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in \overline{S}^j_v\}, \{Y^{i(j+1)}, i \in S \setminus S^j_v\})\\ \approx_{O(jh\epsilon)+O(h\epsilon)+2\epsilon}&(U_\ell, \{R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in \overline{S}^j_v\}, \{Y^{i(j+1)}, i \in S \setminus S^j_v\}).\end{aligned}$$

Since we have already fixed $(Y^{i(j+1)}, i \in S^j_v)$ and $(R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^j_v)$ before, we have that (note that $S^{j+1}_v = S^j_v \cup \overline{S}^j_v$)

$$\begin{aligned}&(R^{v(j+1)}_{\mathsf{Ind}_{v(j+1)}}, \{R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^{j+1}_v\}, \{Y^{i(j+1)}, i \in S\})\\ \approx_{O((j+1)h\epsilon)}&(U_\ell, \{R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^{j+1}_v\}, \{Y^{i(j+1)}, i \in S\}).\end{aligned}$$

Since we have fixed $(\{Y^{i(\leq j)}, i \in S\}, \{R^{i(\leq j-1)}_{\mathsf{Ind}_{i(\leq j-1)}}, i \in S\})$ before, we have that

$$\begin{aligned}&(R^{v(j+1)}_{\mathsf{Ind}_{v(j+1)}}, \{Y^{i(\leq j+1)}, i \in S\}, \{R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^{j+1}_v\}, \{R^{i(\leq j)}_{\mathsf{Ind}_{i(\leq j)}}, i \in S\})\\ \approx_{O((j+1)h\epsilon)}&(U_\ell, \{Y^{i(\leq j+1)}, i \in S\}, \{R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S^{j+1}_v\}, \{R^{i(\leq j)}_{\mathsf{Ind}_{i(\leq j)}}, i \in S\}).\end{aligned}$$

In addition, conditioned on the fixing of $(\{Y^{i(\leq j+1)}, i \in S\}, \{R^{i(\leq j)}_{\mathsf{Ind}_{i(\leq j)}}, i \in S\})$, we have that

1. $X$ and $Y$ are still independent.
2. $(R^{i(j+1)}_{\mathsf{Ind}_{i(j+1)}}, i \in S)$ are all deterministic functions of $X$.
3. The average conditional min-entropy of $X$ is at least $k-jh\ell$ and the average conditional min-entropy of $Y^v$ is at least $0.9k-(j+1)h(h^2+12)\ell$.

Thus the lemma is proved. □

Now we have the following lemma.

**Lemma 5.5.** *Assume that $k \geq 2(bh+2)(h^2+12)\ell$, $X$ is an $(n,k)$-source and $Y$ is an $N \times 0.9k$ SR-source independent of $X$, with $N = 2^d = \text{poly}(n)$ such that there exists a subset $S \subset [N]$ with $|S| > 0.9N$ and for any $i \in S$, $Y^i$ is uniform. Let $Z = Z^1 \circ \cdots \circ Z^N = \mathsf{SSR}(X,Y)$. Then for any subset $S' \subset S$ with $|S'| = h$, we have that*

$$(Z^i, i \in S') \approx_\epsilon U_{h\ell},$$

*where $\epsilon = 2^{-\Omega(\ell)}$.*

*Proof.* We order the elements in $S'$ to be $i_1 < i_2 < \cdots < i_h$. Since $S' \subset S$, for any $j \in [h]$ we have that $Y^{i_j}$ is uniform. We now apply Lemma 5.4 to the set $S'$. Note that $f^b(i) = i - 1$, thus for any $v \in S'$ we have $S'^b_v = \{i \in S' : i < v\}$. Also note that $Z^i = R^{ib}_{\mathsf{Ind}_{ib}}$ for any $i \in [N]$. Thus by Lemma 5.4, for any $j \in [h]$ we have that

$$(Z^{i_j}, Z^{i_1}, \cdots, Z^{i_{j-1}}) \approx_{O(bh2^{-\Omega(\ell)})} (U_\ell, Z^{i_1}, \cdots, Z^{i_{j-1}}).$$

Thus we have that

$$(Z^{i_1}, \cdots, Z^{i_h}) \approx_\epsilon U_{h\ell},$$

where $\epsilon = O(bh^2 2^{-\Omega(\ell)}) = 2^{-\Omega(\ell)}$ since $\ell = k^\beta > k^\alpha$, $h < 2k^\alpha$ and $b < \log n = k^{O(1)}$. □

Now we can describe the lightest bin protocol, defined in [Li13].

**Lightest bin protocol:** Assume there are $N$ strings $\{z^i, i \in [N]\}$ where each $z_i \in \{0,1\}^m$ with $m > \log N$. The output of a lightest bin protocol with $r < N$ bins is a subset $T \subset [N]$ that is obtained as follows. Imagine that each string $z^i$ is associated with a player $P_i$. Now, for each $i$, $P_i$ uses the first $\log r$ bits of $z_i$ to select a bin $j$, i.e., if the first $\log r$ bits of $z_i$ is the binary expression of $j-1$, then $P_i$ selects bin $j$. Now let bin $l$ be the bin that is selected by the fewest number of players. Then

$$T = \{i \in [N] : P_i \text{ selects bin } l.\}$$

To analyze the protocol we first need the following lemma.

**Lemma 5.6.** *For any integer $h$, assume that $\overline{X} = (X_1, \cdots, X_n)$ and $\overline{X'} = (X'_1, \cdots, X'_n)$ are two distributions over $\{0,1\}^n$ such that for any subset $S \subset [n]$ with $|S| = h$, $(X_i, i \in S) \approx_\epsilon (X'_i, i \in S)$. Let $X = \sum_{i=1}^n X_i$ and $X' = \sum_{i=1}^n X'_i$. For any $i \in [n]$, let $\mu_i = E[X_i]$ and $\mu'_i = E[X'_i]$. Let $\mu = \sum_{i=1}^n \mu_i = E[X]$ and $\mu' = \sum_{i=1}^n \mu'_i = E[X']$. Then we have*

$$\left| E[(X-\mu)^h] - E[(X'-\mu')^h] \right| \leq (h+2)n^h\epsilon.$$

*Proof.* We prove the lemma by establishing the following two claims.

**Claim 5.7.**

$$\left| E[(X-\mu)^h] - E[(X'-\mu)^h] \right| \leq 2n^h\epsilon.$$

To show this claim, note that $(X-\mu)^h = (\sum_{i=1}^n (X_i - \mu_i))^h$ is the sum of $n^h$ terms, each of the form $\Pi_{i=1}^n (X_i - \mu_i)^{h_i}$ such that $\forall i, h_i \geq 0$ and $\sum_{i=1}^n h_i = h$. Similarly, $(X'-\mu)^h = (\sum_{i=1}^n (X'_i - \mu_i))^h$ is also the sum of $n^h$ terms, each of the form $\Pi_{i=1}^n (X'_i - \mu_i)^{h_i}$ such that $\sum_{i=1}^n h_i = h$. By linearity of expectation, $E[(X-\mu)^h]$ is the sum of the expectations of these $n^h$ terms. By the triangle inequality,

$|E[(X-\mu)^h]-E[(X'-\mu)^h]|$ is at most the sum of $|E[\Pi_{i=1}^n(X_i-\mu_i)^{h_i}]-E[\Pi_{i=1}^n(X_i'-\mu_i)^{h_i}]|$ for these $n^h$ terms. Now for any fixed $(h_1,\cdots,h_n)$, we consider $|E[\Pi_{i=1}^n(X_i-\mu_i)^{h_i}]-E[\Pi_{i=1}^n(X_i'-\mu_i)^{h_i}]|$.

Since $\sum_{i=1}^n h_i = h$, the term $\Pi_{i=1}^n(X_i-\mu_i)^{h_i}$ involves at most $h$ $X_i$'s. Let $S$ stand for the set of the indexes of the $X_i$'s. Let $X_S = (X_i, i\in S)$ and $X_S' = (X_i', i\in S)$. Thus $s = |S| \le h$ and $X_S \approx_\epsilon X_S'$. We have

$$\begin{aligned}\left|E[\Pi_{i=1}^n(X_i-\mu_i)^{h_i}]-E[\Pi_{i=1}^n(X_i'-\mu_i)^{h_i}]\right| &= \left|\sum_{x\in\{0,1\}^s}(\Pr[X_S=x]-\Pr[X_S'=x])\Pi_{i\in S}(x_i-\mu_i)^{h_i}\right| \\ &\le \sum_{x\in\{0,1\}^s}\left|\Pi_{i\in S}(x_i-\mu_i)^{h_i}\right|\left|\Pr[X_S=x]-\Pr[X_S'=x]\right|.\end{aligned}$$

For any $x\in\{0,1\}^s$, we have $|\Pi_{i\in S}(x_i-\mu_i)^{h_i}|\le 1$. Thus

$$\left|E[\Pi_{i=1}^n(X_i-\mu_i)^{h_i}]-E[\Pi_{i=1}^n(X_i'-\mu_i)^{h_i}]\right| \le \sum_{x\in\{0,1\}^s}\left|\Pr[X_S=x]-\Pr[X_S'=x]\right| \le 2\epsilon.$$

Therefore

$$\left|E[(X-\mu)^h]-E[(X'-\mu)^h]\right| \le 2n^h\epsilon.$$

**Claim 5.8.**

$$\left|E[(X'-\mu)^h]-E[(X'-\mu')^h]\right| \le hn^h\epsilon.$$

To show this claim, note that $E[(X'-\mu)^h] = \sum_{x\in\{0,1\}^n}\Pr[\overline{X'}=x](\sum x_i-\mu)^h$ and $E[(X'-\mu')^h] = \sum_{x\in\{0,1\}^n}\Pr[\overline{X'}=x](\sum x_i-\mu')^h$. Thus

$$\begin{aligned}\left|E[(X'-\mu)^h]-E[(X'-\mu')^h]\right| &= \left|\sum_{x\in\{0,1\}^n}\Pr[\overline{X'}=x]((\sum x_i-\mu)^h-(\sum x_i-\mu')^h)\right| \\ &\le \sum_{x\in\{0,1\}^n}\Pr[\overline{X'}=x]\left|(\sum x_i-\mu)^h-(\sum x_i-\mu')^h\right|.\end{aligned}$$

Now note that $(\sum x_i-\mu)^h = (\sum_{i=1}^n(x_i-\mu_i))^h$ is the sum of $n^h$ terms, each of the form $\Pi_{i=1}^n(x_i-\mu_i)^{h_i}$ such that $\sum_{i=1}^n h_i = h$. Similarly, $(\sum x_i-\mu')^h = (\sum_{i=1}^n(x_i-\mu_i'))^h$ is also the sum of $n^h$ terms, each of the form $\Pi_{i=1}^n(x_i-\mu_i')^{h_i}$ such that $\sum_{i=1}^n h_i = h$. Now for any fixed $(h_1,\cdots,h_n)$, we consider $|\Pi_{i=1}^n(x_i-\mu_i)^{h_i}-\Pi_{i=1}^n(x_i-\mu_i')^{h_i}|$.

Note that for any $i$, $|\mu_i-\mu_i'|\le\epsilon$ since $X_i\approx_\epsilon X_i'$. Thus for any $i$, we have $|(x_i-\mu_i)-(x_i-\mu_i')| = |\mu_i-\mu_i'|\le\epsilon$. We have the following fact.

**Fact 5.9.** *Assume that we have $x_1,x_2,y_1,y_2\in[-1,1]$ such that $|x_1-y_1|\le\epsilon_1$ and $|x_2-y_2|\le\epsilon_2$. Then*

$$|x_1x_2-y_1y_2|\le\epsilon_1+\epsilon_2.$$

Indeed, we have

$$|x_1x_2 - y_1y_2| = |x_1(x_2 - y_2) + y_2(x_1 - y_1)| \leq |x_1(x_2 - y_2)| + |y_2(x_1 - y_1))| \leq \epsilon_1 + \epsilon_2.$$

Using this fact, we thus obtain that

$$|\Pi_{i=1}^n (x_i - \mu_i)^{h_i} - \Pi_{i=1}^n (x_i - \mu_i')^{h_i}| \leq \sum_{i=1}^n h_i\epsilon = h\epsilon.$$

Therefore, for any $x \in \{0,1\}^n$, we have $|(\sum x_i - \mu)^h - (\sum x_i - \mu')^h| \leq hn^h\epsilon$ and thus

$$\begin{aligned}\left|E[(X' - \mu)^h] - E[(X' - \mu')^h]\right| &\leq \sum_{x\in\{0,1\}^n} \Pr[\overline{X'} = x] \left|(\sum x_i - \mu)^h - (\sum x_i - \mu')^h\right| \\ &\leq hn^h\epsilon \sum_{x\in\{0,1\}^n} \Pr[\overline{X'} = x] = hn^h\epsilon.\end{aligned}$$

By the above two claims and the triangle inequality, the lemma is proved. □

We now have the following lemma.

**Lemma 5.10.** *For every constant $0 < \gamma < 1$ there exists a constant $C_1 > 1$ such that the following holds. For any $n, k, m, N \in \mathbb{N}$, any even integer $h \geq C_1$ and any $\epsilon > 0$ with $N \geq h^2$, $\epsilon < N^{-6h}$, $k > 20h(\log n + \log(1/\epsilon))$ and $m > 10(\log n + \log(1/\epsilon))$, assume that we have $N$ sources $\{Z_1^i, i \in [N]\}$ over $m$ bits and a subset $S \subset [N]$ with $|S| \geq \delta N$ for some constant $\delta > 1/2$, such that for any $S' \subset S$ with $|S'| = h$, we have*

$$(Z_1^i, i \in S') \approx_\epsilon U_{hm}.$$

*Let $Z_1 = Z_1^1 \circ \cdots \circ Z_1^N$. Use $Z_1$ to run the lightest bin protocol with $r = \frac{\gamma^2}{16h} N^{1-\frac{2}{\sqrt{h}}}$ bins[1] and let the output contain $N_2$ elements $\{i_1, i_2, \cdots, i_{N_2} \in [N]\}$. Assume that $X$ is an $(n,k)$ source independent of $Z_1$. For any $j \in [N_2]$, let $Z_2^j = \mathsf{Ext}(X, Z_1^{i_j})$ where $\mathsf{Ext}$ is the strong seeded extractor in theorem 3.8 that has seed length $m$ and outputs $m_2 = k/(2h)$ bits with error $\epsilon$. Then with probability at least $1 - N^{-\sqrt{h}/2}$ over the fixing of $Z_1$, there exists a subset $S_2 \subset [N_2]$ with $|S_2| \geq \delta(1-\gamma)N_2$ such that for any $S_2' \subset S_2$ with $|S_2'| = h$, we have*

$$(Z_2^i, i \in S_2') \approx_{\epsilon_2} U_{hm_2}$$

*with $\epsilon_2 < N_2^{-6h}$ and $m_2 > 10(\log n + \log(1/\epsilon_2))$.*

*Proof.* Note that the lightest bin contains $N_2 \leq N/r = \frac{16h}{\gamma^2} N^{\frac{2}{\sqrt{h}}}$ elements. We first show that with high probability every bin contains at least $\delta(1-\gamma)N_2$ elements in $S$.

Consider a particular bin and consider the choices of the $Z_1^i$'s with $i \in S$. Let $s = |S|$. Let $V_i$ be the indicator variable of whether $Z_1^i$ chooses this bin and let $V = \sum_{i\in S} V_i$. Let $p_i = \Pr[V_i = 1]$ and $q_i = \Pr[V_i = 0]$. Then we have

[1] For simplicity, we assume that $r$ is a power of 2. If not, we can always replace it with a power of 2 that is at most $2r$. This does not affect our analysis.

$$E[V] = \sum_{i \in S} E[V_i] = \sum_{i \in S} p_i.$$

We know for any $i \in S$, $Z_1^i$ is $\epsilon$-close to uniform. Thus $\Pr[V_i = 1] \geq 1/r - \epsilon$. Therefore

$$E[V] \geq (1/r - \epsilon)s.$$

Note that

$$\begin{aligned}\Pr[V < 1/r(1-\gamma)s] &\leq \Pr[|V - E[V]| > \gamma s/r - \epsilon s] \\ &\leq \Pr[|V - E[V]| > \gamma s/(2r)],\end{aligned}$$

since $\epsilon s < 1$ and $\gamma$ is a constant $< 1$.

Now since $h$ is an even number, by Markov's inequality we have

$$\Pr[V < 1/r(1-\gamma)s] \leq \Pr[(V - E[V])^h > (\gamma s/(2r))^h] \leq \frac{E[(V - E[V])^h]}{(\gamma s/(2r))^h}.$$

We now estimate $E[(V - E[V])^h]$. First, if the $V_i$'s are truly $h$-wise independent, we have the following claim by Bellare and Rompel [BR94].

**Claim 5.11.** *[BR94] Assume that $(V_1, \cdots, V_{n'})$ are $h$-wise independent random variables over $\{0,1\}^{n'}$. Let $V = \sum_{i=1}^{n'} V_i$ and $\mu = E[V]$. When $h$ is even, we have*

$$E[(V - \mu)^h] \leq 8(\mu h + h^2)^{h/2}.$$

Thus, if the $V_i$'s are truly $h$-wise independent, we have that $\mu = E[V] = s/r \geq \delta N/r = \frac{16\delta h}{\gamma^2} N^{\frac{2}{\sqrt{h}}} > \frac{8h}{\gamma^2} N^{\frac{2}{\sqrt{h}}} > h$. Therefore $\mu h > h^2$ and thus

$$E[(V - \mu)^h] \leq 8(\mu h + h^2)^{h/2} < 8(2\mu h)^{h/2} = 8(2sh/r)^{h/2}.$$

Now, since the $V_i$'s are actually $h$-wise $\epsilon$-close to being independent, by Lemma 5.6 we have that

$$E[(V - E[V])^h] < 8(2sh/r)^{h/2} + (h+2)s^h\epsilon < 9(2sh/r)^{h/2},$$

since $s \leq N$ and $\epsilon < N^{-6h}$.

Thus, we have that

$$\Pr[V < 1/r(1-\gamma)s] \leq \frac{E[(V - E[V])^h]}{(\gamma s/(2r))^h} < \frac{9(2sh/r)^{h/2}}{(\gamma s/(2r))^h} = 9\left(\frac{8rh}{\gamma^2 s}\right)^{\frac{h}{2}}$$

Note that $s = \delta N$ with $\delta > 1/2$ and $r = \frac{\gamma^2}{16h} N^{1 - \frac{2}{\sqrt{h}}}$, we thus have

$$\Pr[V < 1/r(1-\gamma)s] < 9\left(\frac{8rh}{\gamma^2 s}\right)^{\frac{h}{2}} < 9(N^{-\frac{2}{\sqrt{h}}})^{\frac{h}{2}} = 9N^{-\sqrt{h}}.$$

Thus by the union bound, we have that the probability that every bin contains at least $1/r(1-\gamma)s$ elements in $S$ is at least $1-9rN^{-\sqrt{h}} = 1-\frac{9\gamma^2}{16h}N^{1-\frac{2}{\sqrt{h}}-\sqrt{h}} > 1-\frac{1}{2}N^{-\sqrt{h}/2}$. When this happens, let $S_2$ be the set of elements in $S$ in the lightest bin. Then we have $|S_2| \geq 1/r(1-\gamma)s \geq \delta(1-\gamma)N_2$.

Next, we show that with high probability the new sources with indexes in $S_2$ are $h$-wise close to uniform. For this, consider any $S' \subset S$ with $|S'| = h$. Let $S' = \{i_1, \cdots, i_h\}$ and for any $j \in [h]$, let $W^{i_j} = \mathsf{Ext}(X, Z_1^{i_j})$. Note that $(Z_1^{i_1}, \cdots, Z_1^{i_h}) \approx_\epsilon U_{hm}$. First assume that $(Z_1^{i_1}, \cdots, Z_1^{i_h})$ is indeed uniform. Thus by Theorem 3.8 we have

$$(W^{i_1}, Z_1^{i_1}) \approx_\epsilon (U_{m_2}, Z_1^{i_1}).$$

Next, note that for any $j \in [h]$ with $j \geq 2$, conditioned on the fixing of $(Z_1^{i_1}, \cdots, Z_1^{i_{(j-1)}})$, we have that $Z_1^{i_j}$ is still uniform. Moreover, conditioned on this fixing, $(W^{i_1}, \cdots, W^{i_{(j-1)}})$ is a deterministic function of $X$. Thus we can further fix $(W^{i_1}, \cdots, W^{i_{(j-1)}})$ and conditioned on this fixing, $Z_1^{i_j}$ is still independent of $X$. Moreover, the average conditional min-entropy of $X$ is at least $k-(j-1)m_2 > k-hm_2 = k/2$. Thus by Theorem 3.8 and Lemma 3.7 (notice that $k$ and $m$ are much bigger than $\log n + \log(1/\epsilon)$) we have

$$(W^{i_j}, Z_1^{i_j}, W^{i_1}, Z_1^{i_1}, \cdots, W^{i_{(j-1)}}, Z_1^{i_{(j-1)}}) \approx_{2\epsilon} (U_{m_2}, Z_1^{i_j}, W^{i_1}, Z_1^{i_1}, \cdots, W^{i_{(j-1)}}, Z_1^{i_{(j-1)}}).$$

Note that for any $j \in [h]$, conditioned on the fixing of $Z_1^{i_j}$, $W^{i_j}$ is a deterministic function of $X$ and is thus independent of all $\{Z_1^{i_j}\}$'s. Thus we have that

$$(W^{i_1}, Z_1^{i_1}, \cdots, Z_1^{i_h}) \approx_\epsilon (U_{m_2}, Z_1^{i_1}, \cdots, Z_1^{i_h})$$

and for any $j \geq 2$,

$$(W^{i_j}, W^{i_1}, \cdots, W^{i_{(j-1)}}, Z_1^{i_1}, \cdots, Z_1^{i_h}) \approx_{2\epsilon} (U_{m_2}, W^{i_1}, \cdots, W^{i_{(j-1)}}, Z_1^{i_1}, \cdots, Z_1^{i_h}).$$

This implies that

$$(W^{i_1}, \cdots, W^{i_h}, Z_1^{i_1}, \cdots, Z_1^{i_h}) \approx_{(2h-1)\epsilon} (U_{hm_2}, Z_1^{i_1}, \cdots, Z_1^{i_h}).$$

Adding back the error where $(Z_1^{i_1}, \cdots, Z_1^{i_h}) \approx_\epsilon U_{hm}$, we have

$$(W^{i_1}, \cdots, W^{i_h}, Z_1^{i_1}, \cdots, Z_1^{i_h}) \approx_{2h\epsilon} (U_{hm_2}, Z_1^{i_1}, \cdots, Z_1^{i_h}).$$

Therefore, with probability $1-2N^{-2h}$ over the fixing of $(Z_1^{i_1}, \cdots, Z_1^{i_h})$, we have that $(W^{i_1}, \cdots, W^{i_h})$ is $N^{2h}h\epsilon$-close to uniform. Thus by the union bound (and noticing that $s \leq N$), we have that with probability at least $1-2N^{-h}$ over the fixing of $Z_1$, for any $S' \subset S$ with $|S'| = h$, $(W^i, i \in S')$ is $\epsilon_2 = N^{2h}h\epsilon$-close to uniform. In particular, this implies that the new sources with indexes in $S_2$ are $h$-wise close to uniform.

Note that $\epsilon < N^{-6h}$. Thus $N^{2h}h\epsilon < N^{-4h}h$. Since $N \geq h^2$ we have that $N^{-4h+12\sqrt{h}} < h^{-7h} < h^{-1}(\frac{16h}{\gamma^2})^{-6h}$ for sufficiently large $h$. Thus $\epsilon_2 = N^{2h}h\epsilon < N^{-4h}h < (\frac{16h}{\gamma^2})^{-6h}N^{-12\sqrt{h}} \leq N_2^{-6h}$ since $N_2 \leq N/r = \frac{16h}{\gamma^2}N^{\frac{2}{\sqrt{h}}}$. Also note that $m_2 = k/(2h) > 10(\log n + \log(1/\epsilon)) > 10(\log n + \log(1/\epsilon_2))$. Note that $2N^{-h} < \frac{1}{2}N^{-\sqrt{h}/2}$. By the union bound, the lemma is proved. □

We can now present our construction of extractors for independent sources.

**Algorithm 5.12** (Independent Source Extractor $\mathsf{IExt}$)**.**

**Input:** $X_1, X_2, \cdots$ — independent $(n,k)$-sources with $k \geq \text{polylog}(n)$.
**Output:** $W$ — a random variable close to uniform.

**Sub-Routines and Parameters:**
Let $\mathsf{SR}$ be the function in algorithm 5.1. Let $\mathsf{SSR}$ be the function in algorithm 5.3. Let $\mathsf{SRExt}$ be the extractor in Theorem 3.9. Let $\mathsf{Ext}$ be the strong extractor in theorem 3.8. Let $\mathsf{C}$ be the constant in Lemma 5.2. Let $0 < \alpha, \beta, \gamma < 1$ be three constants with $\alpha < \beta$, to be chosen later.

1. Take $\mathsf{C}$ independent sources $(X_1, \cdots, X_{\mathsf{C}})$ and compute $Y = \mathsf{SR}(X_1, \cdots, X_{\mathsf{C}})$.
2. Take another independent source $X_{\mathsf{C}+1}$ and compute $Z_1 = Z_1^1 \circ \cdots \circ Z_1^{N_1} = \mathsf{SSR}(X_{\mathsf{C}+1}, Y)$ with $N_1 = \text{poly}(n)$, using parameters $\alpha$ and $\beta$.
3. Let $h$ be the parameter in algorithm 5.3 with $k^\alpha \leq h < 2k^\alpha$. Set $t = 1$. While $N_t$ (the number of rows in $Z_t$) is bigger than $h^4$ do the following:
   (a) Run the lightest bin protocol with $Z_t$ and $r_t = \frac{\gamma^2}{16h} N_t^{1-\frac{2}{\sqrt{h}}}$ bins and let the output contain $N_{t+1}$ elements $\{i_1, i_2, \cdots, i_{N_{t+1}} \in [N_t]\}$.
   (b) Take a fresh independent $(n,k)$ source $X_{\mathsf{C}+t+1}$ and for any $j \in [N_{t+1}]$, compute $Z_{t+1}^j = \mathsf{Ext}(X_{\mathsf{C}+t+1}, Z_t^{i_j})$ and output $m_2 = k/(2h)$ bits.
   (c) Let $Z_{t+1} = Z_{t+1}^1 \circ \cdots \circ Z_{t+1}^{N_{t+1}}$. Set $t = t + 1$.
4. At the end of the above iteration we get a source $Z_t$ with at most $h^4$ rows. Take another $c_2$ independent $(n,k)$ sources $X_{\mathsf{C}+t+1}, \cdots, X_{\mathsf{C}+t+c_2}$. The final output is $W = \mathsf{SRExt}(Z_t, X_{\mathsf{C}+t+1}, \cdots, X_{\mathsf{C}+t+c_2})$.

We now have the following theorem.

**Theorem 5.13.** *For every constant $\eta > 0$ there exists a constant $C_0 > 1$ such that for any $n, k \in \mathbb{N}$ with $n \geq C_0$ and $k \geq \log^{2+\eta} n$, the above construction is an extractor that uses $O\left(\frac{1}{\eta}\right) + O(1)$ independent $(n,k)$ sources and outputs $m = \Omega(k)$ bits that are $1/\text{poly}(n) + 2^{-k^{\Omega(1)}}$-close to uniform.*

*Proof.* By Lemma 5.2, $Y = \mathsf{SR}(X_1, \cdots, X_{\mathsf{C}})$ is $1/\text{poly}(n)$-close to another source $Y' = Y'^1 \circ \cdots \circ Y'^N$ where $N = \text{poly}(n)$ and each $Y'^i$ has $0.9k$ bits, such that there exists a subset $S \subset [N]$ with $|S| > 0.9N$ and for any $i \in S$, $Y'^i$ is uniform. We will now proceed as if $Y$ is the source $Y'$. This only adds $1/\text{poly}(n)$ to the final error.

We now want to apply Lemma 5.5 and Lemma 5.10. Before we do this, let us first set the parameters $\alpha$ and $\beta$. Note that $k^\alpha \leq h < 2k^\alpha$, $\ell = k^\beta$ and $b < \log n$. To apply Lemma 5.5, we need that $k \geq 2(bh+2)(h^2+12)\ell$. To apply Lemma 5.10 for the first time, we need that $\epsilon = 2^{-\Omega(\ell)} < N^{-6h}$, $k > 20h(\log n + \log(1/\epsilon))$ and $m = \ell > 10(\log n + \log(1/\epsilon))$ (note that we

only go into the iteration of the lightest bin protocol if $N > h^4$). Altogether, it suffices to have $0 < \alpha < \beta < 1$ satisfy the following conditions.

$$k \geq 3 \log n h^3 \ell,\ 2^{-\Omega(\ell)} < N^{-6h} \text{ and } \ell > 10(\log n + \log(1/\epsilon)).$$

These conditions are satisfied if the following conditions are satisfied.

$$k \geq 24 k^{3\alpha+\beta} \log n \text{ and } \ell = k^\beta \geq C k^\alpha \log n$$

for some constant $C > 1$.

Thus when $k \geq \log^{2+\eta} n$, we can choose $\alpha = \frac{\eta}{16(2+\eta)}$ and $\beta = \frac{16+5\eta}{16(2+\eta)}$. Now note that $Z_1 = Z_1^1 \circ \cdots \circ Z_1^{N_1} = \mathsf{SSR}(X_{\mathsf{C}+1}, Y)$ satisfies the conditions of Lemma 5.10. Thus we can apply Lemma 5.10.

Note that the lightest bin protocol stops only if the number of rows in $Z^t$ is at most $h^4$. Thus before the iteration stops, we always have $N_t > h^4 > h^2$. Thus by Lemma 5.10 the probability of the "bad event" in each iteration is at most $N_t^{-\sqrt{h}/2} < (h^4)^{-\sqrt{h}/2} = h^{-2\sqrt{h}} = 2^{-k^{\Omega(1)}}$. We now compute the number of iterations needed to decrease the number of rows from $N = \text{poly}(n)$ to $h^4$.

In each iteration the number of rows in $Z_t$ decreases from $N_t$ to $N_{t+1} \leq \frac{16h}{\gamma^2} N_t^{\frac{2}{\sqrt{h}}}$. When $N_t \geq h^{\sqrt{h}}$, we have that $N_t^{\frac{2}{\sqrt{h}}} \geq h^2 > \frac{16h}{\gamma^2}$. Thus

$$N_{t+1} \leq \frac{16h}{\gamma^2} N_t^{\frac{2}{\sqrt{h}}} < N_t^{\frac{4}{\sqrt{h}}}.$$

Therefore, as long as $N_t \geq h^{\sqrt{h}}$, in each iteration the number of rows in $Z^t$ decreases from $N_t$ to $N_{t+1} \leq N_t^{\frac{4}{\sqrt{h}}}$. Since initially we have $N_1 = \text{poly}(n)$, the number of iterations needed to decrease the number of rows from $N = \text{poly}(n)$ to $h^{\sqrt{h}}$ is at most

$$\log_{\frac{\sqrt{h}}{4}} \frac{\log N}{\sqrt{h} \log h} = \frac{\log\log N - \frac{1}{2}\log h - \log\log h}{\frac{1}{2}\log h - 2} = O\left(\frac{\log\log n}{\alpha \log k}\right) = O\left(\frac{1}{\alpha}\right) = O\left(\frac{1}{\eta}\right) + O(1).$$

When $\eta < 1$, this number is $O\left(\frac{1}{\eta}\right)$. When $\eta \geq 1$, it is $O(1)$. In either case, as long as $\eta > 0$ is a constant, it is a constant $c'$.

Once $N_t \leq h^{\sqrt{h}}$, in the next iteration we have

$$N_{t+1} \leq \frac{16h}{\gamma^2} N_t^{\frac{2}{\sqrt{h}}} \leq \frac{16h}{\gamma^2} h^2 < h^4.$$

Thus the number of iterations needed to decrease the number of rows from $N = \text{poly}(n)$ to $h^4$ is at most $c_3 = c' + 1$, which is also a constant. Therefore, we can set $\gamma = \frac{1}{5c_3} = \Omega(\frac{\eta}{1+\eta})$. This ensures that in each $N_t$, the fraction of "good rows" is at least $0.9(1-\gamma)^{c_3} > 0.9(1 - c_3\gamma) \geq 0.8 \cdot 0.9 > 1/2$, which satisfies the requirement of Lemma 5.10. Also note that there exists a constant $C_0 = C_0(\eta)$ such that whenever $n \geq C_0$ and $k \geq \log^2 n$ we have $h \geq k^\alpha \geq C_1$ where $C_1$ is the constant in Lemma 5.10. Thus we are all good. Note that the number of independent sources used in the iteration is also $c_3$.

Finally, since when the iteration stops $Z^t$ has at most $h^4 < 16k^{4\alpha} < 16k^{1/4}$ rows, by Theorem 3.9 it suffices to take another constant $c_2$ number of independent sources[2] and output $W = \mathsf{SRExt}(Z_t, X_{\mathsf{C}+c_3+1}, \cdots, X_{\mathsf{C}+c_3+c_2})$, which is $2^{-k^{\Omega(1)}}$-close to uniform. Therefore, altogether the extractor uses $\mathsf{C} + c_3 + c_2 = O\left(\frac{1}{\eta}\right) + O(1)$ independent $(n, k)$ sources, and the error of the extractor is $1/\mathrm{poly}(n) + c_3 2^{-k^{\Omega(1)}} + 2^{-k^{\Omega(1)}} = 1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$. ■

**Remark 5.14.** The condenser based on lightest bin protocol is highly efficient by making use of the property of $k^\alpha$-wise independent. In fact, when $k \geq \log^c n$ for some big enough constant $c > 2$, we will have $N < h^{\sqrt{h}}$ in the first place and thus the lightest bin protocol only takes one iteration. On the other hand, as $c$ approaches 2, the number of iterations increases.

Note that when $n < C_0$, the extractor can be constructed in constant time just by exhaustive search (in fact, we can get a two-source extractor in this way). Thus, we have the following theorem.

**Theorem 5.15.** *For every constant $\eta > 0$ and all $n, k \in \mathbb{N}$ with $k \geq \log^{2+\eta} n$, there is an explicit extractor that uses $O\left(\frac{1}{\eta}\right) + O(1)$ independent $(n, k)$ sources and outputs $m = \Omega(k)$ bits with error $1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$.*

We also have the following two corollaries.

**Corollary 5.16.** *For all $n, k \in \mathbb{N}$ with $k \geq \log^3 n$, there is an explicit extractor that uses $O(1)$ independent $(n, k)$ sources and outputs $m = \Omega(k)$ bits with error $1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$.*

**Corollary 5.17.** *For every constant $0 < \eta < 1$ and all $n, k \in \mathbb{N}$ with $k \geq \log^{2+\eta} n$, there is an explicit extractor that uses $O\left(\frac{1}{\eta}\right)$ independent $(n, k)$ sources and outputs $m = \Omega(k)$ bits with error $1/\mathrm{poly}(n) + 2^{-k^{\Omega(1)}}$.*

## 6 Conclusions and Open Problems

In this paper we construct explicit extractors for a constant number of independent weak random sources with poly-logarithmic min-entropy. This dramatically improves all previous results and brings the construction of independent source extractors close to optimal. The main technical ingredient is the condenser for structured somewhere random sources, based on the connection between condensing somewhere random sources and leader election, as suggested in [Li13]. The condenser makes use of the property that a large fraction of the rows in the somewhere random source are close to $h$-wise independent and uniform, and uses Feige's lightest bin protocol to reduce the number of rows in the somewhere random source. By achieving $h = k^\alpha$ for some constant $0 < \alpha < 1$, the condenser in this paper is highly efficient and thus we obtain our extractor for a constant number of independent sources, even with poly-logarithmic min-entropy.

Several natural open problems remain. First of all, how far can we push this construction? Our ultimate goal is to construct two-source extractors for logarithmic min-entropy. Can we reduce the number of independent sources used in our extractor from a constant number to two? There seems to be some obstacles in the way. For example, in our construction we need to first use a constant number of sources to reduce the error and obtain a source that is $1/\mathrm{poly}(n)$-close to a somewhere

[2] In fact, it suffices to take $c_2 = 2$.

random source which has a large fraction of truly uniform rows. Even using seeded extractors with optimal parameters, this already requires at least three independent sources. However, we hope that this constraint can be removed by combining other techniques used in various constructions of extractors. More realistically, as a first step we may hope to construct two-source extractors for arbitrarily linear min-entropy.

Second, it would be nice to achieve better error. Currently the error of our extractor is $1/\text{poly}(n) + 2^{-k^{\Omega(1)}}$ and ideally we would want to remove the $1/\text{poly}(n)$ term and achieve error $2^{-k^{\Omega(1)}}$. This kind of small error is important for cryptographic applications. We believe this should be possible.

Finally, can we apply the powerful techniques of condensing structured somewhere random sources to constructing extractors and dispersers for other classes of sources, and more generally, to constructing other pseudorandom objects? Since independent source extractors have been used in the constructions of extractors and dispersers for various classes of sources (e.g., affine sources and small space sources), this seems promising.

## References


[BIW04] Boaz Barak, R. Impagliazzo, and Avi Wigderson. Extracting randomness using few independent sources. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 384–393, 2004.

[BKS+05] Boaz Barak, Guy Kindler, Ronen Shaltiel, Benny Sudakov, and Avi Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 1–10, 2005.

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

[BR94] Mihir Bellare and John Rompel. Randomness-efficient oblivious sampling. In *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science*, 1994.

[BRSW06] 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. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, 2006.

[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.

[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.

[DW08] Zeev Dvir and Avi Wigderson. Kakeya sets, new mergers and old extractors. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, 2008.

[DW09] Yevgeniy Dodis and Daniel Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 601–610, 2009.

[Fei99] Uriel Feige. Noncryptographic selection protocols. In IEEE, editor, *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 142–152. IEEE Computer Society Press, 1999.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. *Journal of the 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*, 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*, 2008.

[Li11] Xin Li. Improved constructions of three source extractors. In *Proceedings of the 26th Annual IEEE Conference on Computational Complexity*, 2011.

[Li12a] Xin Li. Non-malleable condensers for arbitrary min-entropy, and almost optimal protocols for privacy amplification. Technical report, Arxiv, 2012. arXiv:1211.0651.

[Li12b] Xin Li. Non-malleable extractors, two-source extractors and privacy amplification. In *Proceedings of the 53nd Annual IEEE Symposium on Foundations of Computer Science*, 2012.

[Li13] Xin Li. New independent source extractors with exponential improvement. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, 2013.

[LRVW03] C. J. Lu, Omer Reingold, Salil Vadhan, and Avi Wigderson. Extractors: Optimal up to constant factors. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 602–611, 2003.

[MW97] Ueli M. Maurer and Stefan Wolf. Privacy amplification secure against active adversaries. In *CRYPTO '97*, 1997.

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

[Rao06] Anup Rao. Extractors for a constant number of polynomially small min-entropy independent sources. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, 2006.

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



ECCC ISSN 1433-8092

http://eccc.hpi-web.de