Electronic Colloquium on Computational Complexity, Report No. 115 (2016)

# Improved Non-Malleable Extractors, Non-Malleable Codes and Independent Source Extractors

Xin Li *
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, U.S.A.
lixints@cs.jhu.edu

**Abstract**

In this paper we give improved constructions of several central objects in the literature of randomness extraction and tamper-resilient cryptography. Our main results are:

(1) An explicit seeded non-malleable extractor with error $\epsilon$ and seed length $d = O(\log n) + O(\log(1/\epsilon)\log\log(1/\epsilon))$, that supports min-entropy $k = \Omega(d)$ and outputs $\Omega(k)$ bits. Combined with the protocol in [DW09], this gives a two round privacy amplification protocol with optimal entropy loss in the presence of an active adversary, for all security parameters up to $\Omega(k/\log k)$, where $k$ is the min-entropy of the shared weak random source. Previously, the best known seeded non-malleable extractors require seed length and min-entropy $O(\log n) + \log(1/\epsilon)2^{O\sqrt{\log\log(1/\epsilon)}}$ [CL16, Coh16a], and only give two round privacy amplification protocols with optimal entropy loss for security parameter up to $k/2^{O(\sqrt{\log k})}$.

(2) An explicit non-malleable two-source extractor for min-entropy $k \geq (1-\gamma)n$, some constant $\gamma > 0$, that outputs $\Omega(k)$ bits with error $2^{-\Omega(n/\log n)}$. We further show that we can efficiently uniformly sample from the pre-image of any output of the extractor. Combined with the connection in [CG14b] this gives a non-malleable code in the two-split-state model with relative rate $\Omega(1/\log n)$. This exponentially improves previous constructions, all of which only achieve rate $n^{-\Omega(1)}$.[1]

(3) Combined with the techniques in [BADTS16], our non-malleable extractors give a two-source extractor for min-entropy $O(\log n \log\log n)$, which also implies a $K$-Ramsey graph on $N$ vertices with $K = (\log N)^{O(\log\log\log N)}$. Previously the best known two-source extractor in [BADTS16] requires min-entropy $\log n 2^{O(\sqrt{\log n})}$, which gives a Ramsey graph with $K = (\log N)^{2^{O(\sqrt{\log\log\log N})}}$. We further show a way to reduce the problem of constructing seeded $s$-source non-malleable extractors to the problem of constructing non-malleable $(s+1)$-source extractors. Using the non-malleable 10-source extractor with optimal error in [CZ14], we obtain a seeded non-malleable 9-source extractor with optimal seed length, which in turn gives a 10-source extractor for min-entropy $O(\log n)$. Previously the best known extractor for such min-entropy requires $O(\log\log n)$ sources [CS16].

Independent of our work, Cohen [Coh16d] obtained similar results to (1) and the two-source extractor, except the dependence on $\epsilon$ is $\log(1/\epsilon)\text{polylog}\log(1/\epsilon)$ and the two-source extractor requires min-entropy $\log n\text{polylog}\log n$.

*Supported in part by NSF Grant CCF-1617713.

[1] The work of Aggarwal et. al [ADKO15] had a construction which "achieves" constant rate, but recently the author found an error in their proof.

ISSN 1433-8092

# 1 Introduction

Randomness extractors are fundamental objects in the study of pseudorandomness, a branch of modern theoretical computer science. Their motivations come from the need of uniform random bits in many applications, such as randomized algorithms, distributed computing, and cryptography, and the fact that natural random sources are almost always biased. Informally, randomness extractors transform imperfect random sources (whether naturally so or as a result of adversarial information leakage) into nearly uniform random bits, which can then be used in standard applications. Over the past decades randomness extractors have been extensively studied.

To model imperfect randomness, we use the by now standard model of a general weak random source with a certain amount of entropy.

**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* $H_\infty(X)/n$.

It is well known that by just having one weak source as input, no deterministic extractor can work for all $(n,k)$ sources even if $k = n-1$. Several ways are thus explored to get around this. One approach, introduced by Nisan and Zuckerman [NZ96], is to give the extractor an additional independent short uniform random seed. This results in the so called *seeded extractors*.

**Definition 1.2.** (Seeded Extractor) A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a $(k,\varepsilon)$-*extractor* if for every source $X$ with min-entropy $k$ and independent $Y$ which is uniform on $\{0,1\}^d$,

$$|\mathsf{Ext}(X,Y) - U_m| \leq \epsilon.$$

If in addition we have $|(\mathsf{Ext}(X,Y),Y) - (U_m,Y)| \leq \epsilon$ then we say it is a *strong* $(k,\varepsilon)$-*extractor*.

One can show that seeded extractors with very good parameters exist for all $(n,k)$ sources, and with a long line of research their constructions are now close to optimal (e.g., [LRVW03, GUV09, DW08, DKSS09]). Besides their original motivation, seeded extractors have found many other applications in theoretical computer science.

This paper, on the other hand, focuses on several other kinds of randomness extractors which have gained a lot of attention recently. The first one is extractors for *independent sources*. Here, the extractor does not have any additional uniform random seed, but instead it is given as input more than one independent general weak random sources. The probabilistic method shows that deterministic extractors exist for just two independent $(n,k)$ sources with $k \geq \log n + O(1)$. In fact, with high probability a random function is such a two-source extractor. However, giving explicit constructions of such extractors turns out to be quite challenging.

The second kind of extractors we study here, focuses on the case where either the seed or the source is tampered with by an adversary. In this case, one useful and natural property to impose on the extractors is to ensure that the non-tampered output of the extractor is (close to) uniform even given the tampered output. This leads to a large class of generalized randomness extractors called *non-malleable extractors*.

**Definition 1.3** (Tampering Funtion)**.** For any function $f : S \to S$, $f$ has a fixed point at $s \in S$ if $f(s) = s$. We say $f$ has no fixed points in $T \subseteq S$, if $f(t) \neq t$ for all $t \in T$. We say $f$ has no fixed points if $f(s) \neq s$ for all $s \in S$.

When the tampering acts on the seed of a seeded extractor, one obtains a generalization of strong seeded extractors called *seeded non-malleable extractors*, originally introduced by Dodis and Wichs in [DW09].

**Definition 1.4** (Non-malleable extractor)**.** A function $\text{snmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a seeded non-malleable extractor for min-entropy $k$ and error $\epsilon$ if the following holds : If $X$ is a source on $\{0,1\}^n$ with min-entropy $k$ and $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ is an arbitrary tampering function with no fixed points, then

$$|\text{snmExt}(X, U_d) \circ \text{snmExt}(X, \mathcal{A}(U_d)) \circ U_d - U_m \circ \text{snmExt}(X, \mathcal{A}(U_d)) \circ U_d| < \epsilon$$

where $U_m$ is independent of $U_d$ and $X$.

When the tampering acts on the sources in an independent source extractor, one obtains a generalization of independent source extractors called seedless non-malleable extractors, originally introduced by Cheraghchi and Guruswami [CG14b].

**Definition 1.5** (Seedless Non-Malleable $C$-Source Extractor)**.** A function $\mathsf{nmExt} : (\{0,1\}^n)^C \to \{0,1\}^m$ is a $(k, \epsilon)$-seedless non-malleable extractor for $C$ independent sources, if it satisfies the following property: Let $X_1, \cdots, X_C$ be $C$ independent $(n, k)$ sources, and $f_1, \cdots, f_C : \{0,1\}^n \to \{0,1\}^n$ be $C$ arbitrary tampering functions such that there exists an $f_i$ with no fixed points, then

$$|\mathsf{nmExt}(X_1, \cdots, X_C) \circ \mathsf{nmExt}(f_1(X_1), \cdots, f_C(X_2)) - U_m \circ \mathsf{nmExt}(f_1(X_1), \cdots, f_C(X_2))| < \epsilon.$$

Further, we say that the non-malleable extractor is strong if for *every* $i$, we have that

$$|\mathsf{nmExt}(X_1, \cdots, X_C) \circ \mathsf{nmExt}(f_1(X_1), \cdots, f_C(X_2)) \circ X_i - U_m \circ \mathsf{nmExt}(f_1(X_1), \cdots, f_C(X_2)) \circ X_i| < \epsilon.$$

We can also generalize the definition to handle more than one tampering functions.

**Definition 1.6** (Seeded $t$-Non-malleable extractor)**.** A function $\text{snmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a seeded $t$-non-malleable extractor for min-entropy $k$ and error $\epsilon$ if the following holds : If $X$ is a source on $\{0,1\}^n$ with min-entropy $k$ and $\mathcal{A}_1, \cdots, \mathcal{A}_t : \{0,1\}^d \to \{0,1\}^d$ are $t$ arbitrary tampering functions with no fixed points, then

$$|\text{snmExt}(X, U_d) \circ \{\text{snmExt}(X, \mathcal{A}_i(U_d)), i \in [t]\} \circ U_d - U_m \circ \{\text{snmExt}(X, \mathcal{A}_i(U_d)), i \in [t]\} \circ U_d| < \epsilon$$

where $U_m$ is independent of $U_d$ and $X$.

This definition can also be generalized to the case of seeded $t$-non-malleable extractor for more than one weak sources in the obvious way, and we omit the definition here.

As stated above, seeded non-malleable extractors were first introduced by Dodis and Wichs in [DW09], to study a cryptographic problem known as *privacy amplification*. Although they seem to be irrelevant to independent source extractors, it turns out that these two kinds of extractors are closely related. Indeed, since the author's previous work [Li12b, Li13b] which first established connections between seeded non-malleable extractors and independent source extractors, their connections have been demonstrated in several subsequent work. In particular, with other techniques, these connections have led to the recent breakthrough construction of two source extractors by Chattopadhyay and Zuckerman [CZ16]. We now briefly review previous work below.

**Independent source extractors.** The introduction of independent source extractors, as well as the first explicit construction of a two-source extractor appeared in [CG88], where Chor and Goldreich showed that the well known Lindsey's lemma gives an extractor for two independent $(n, k)$ sources with $k > n/2$. Since then there has been essentially no progress until Barak et. al[BIW04] introduced new techniques in additive combinatorics into this problem, and constructed extractors for $O(1/\delta)$ independent $(n, \delta n)$ sources. Subsequently, a long line of fruitful results [BIW04, BKS+05, Raz05, Bou05, Rao06, BRSW06, Li11, Li13b, Li13a, Li15b, Coh15] has introduced many new techniques and culminated in the three source extractor of exponentially small error for poly-logarithmic min-entropy by the author [Li15b]. In the case of two-source extractors, Bourgain [Bou05] gave a construction that breaks the entropy rate 1/2 barrier, and works for two independent $(n, 0.49n)$ sources. In a different work, Raz [Raz05] gave an incomparable result of two source extractors which requires one source to have min-entropy larger than $n/2$, while the other source can have min-entropy $O(\log n)$. In a recent result, Chattopadhyay and Zuckerman [CZ16] greatly improved the situation and gave the first explicit two-source extractor for $(n, k)$ sources with $k \geq \log^C n$ for some large enough constant $C$. Their construction only outputs one bit but this was later improved by the author to output almost all entropy [Li16] and by Meka [Mek15] to work for smaller min-entropy.

Very recently, there has been a new line of work focusing on constructing explicit independent source extractors for very small min-entropy (i.e., near logarithmic). Cohen and Schulman [CS16] constructed extractors for $O(1/\delta)$ sources with min-entropy $\log^{1+\delta} n$. Chattopadhyay and Li [CL16] improved this result to give an explicit extractor for $O(1)$ sources with min-entropy $\log n 2^{O(\sqrt{\log \log n})}$, and this was subsequently improved by Cohen [Coh16a] to achieve a 5-source extractor with the same entropy requirement. Finally, Ben-Aroya et. al [BADTS16] further improves this and achieves a two-source extractor for min-entropy $\log n 2^{O(\sqrt{\log \log n})}$.

**Seeded non-malleable extractors and privacy amplification.** As mentioned above, seeded non-malleable extractors were first introduced by Dodis and Wichs [DW09] to study the question of privacy amplification with an active adversary, and they were later found to have close connections to independent source extractors. Thus, any progress in non-malleable extractors is likely to lead to progress in both the privacy amplification problem and the independent source extractor problem.

Privacy amplification [BBR88] is a basic problem in information theoretic cryptography, where two parties with local (non-shared) uniform random bits communicate through a public channel to convert a shared secret weak random source $\mathbf{X}$ into shared secret nearly uniform random bits. The communication channel is watched by an adversary Eve, who has unlimited computational power and tries to corrupt the protocol. Standard strong seeded extractors are enough to give very efficient protocols for this problem in the case where Eve is passive (i.e., can only see the messages but cannot change them). In the more complicated case where Eve is active (i.e., can arbitrarily change, delete and reorder messages), the goal is to design a protocol that uses as few number of interactions as possible, and outputs a shared uniform random string $\mathbf{R}$ as long as possible (the difference between the length of the output and $H_\infty(\mathbf{X})$ is called *entropy loss*). The protocol is associated with a security parameter $s$, and ensures that if Eve is active, then the probability that Eve can successfully make the two parties output two different strings without being detected is at most $2^{-s}$. On the other hand, if Eve remains passive, then the two parties should achieve shared secret random bits that are $2^{-s}$-close to uniform. We refer the readers to [DLWZ14] for a formal definition.

Much research has been devoted to this problem [MW97, DKRS06, DW09, RW03, KR09, CKOR10, DLWZ14, CRS14, Li12a, Li12b, Li15a]. It is known that when the entropy rate of $\mathbf{X}$ is large, i.e., bigger than $1/2$, there exist protocols that take only one round (e.g., [MW97, DKRS06]), albeit with quite large entropy loss. When the entropy rate of $\mathbf{X}$ is smaller than $1/2$, [DW09] showed that any protocol has to take at least two rounds with entropy loss at least $O(s)$. Thus, the natural goal is to design a two-round protocol with such optimal entropy loss, for any possible security parameter (ideally up to $\Omega(k)$). However, all protocols before the work of [DLWZ14] require $O(s)$ rounds or entropy loss $O(s^2)$.

In [DW09], Dodis and Wichs further showed that two-round privacy amplification protocols with optimal entropy loss can be constructed using explicit seeded non-malleable extractors. Using the probabilistic method, they showed the existence of non-malleable extractors when $k > 2m + 2\log(1/\varepsilon) + \log d + 6$ and $d > \log(n - k + 1) + 2\log(1/\varepsilon) + 5$. However, they were not able to give any explicit construction. The first explicit seeded non-malleable extractor was constructed in [DLWZ14], with subsequent improvements in [CRS14, Li12a, DY13, Li12b]. Unfortunately all these constructions require min-entropy at least $0.49n$, and thus only give two-round privacy amplification protocols with optimal entropy loss for such min-entropy. Although, combined with other ideas, [DLWZ14] also gives poly$(1/\delta)$ round protocols with optimal entropy loss for min-entropy $k \geq \delta n$, any constant $\delta > 0$. Subsequently, without improving on the non-malleable extractors, the author [Li12b] gave a two-round protocol with optimal entropy loss for min-entropy $k \geq \delta n$, any constant $\delta > 0$. Using a relaxation of non-malleable extractors called non-malleable condensers, the author [Li15a] also obtained a two-round protocol with optimal entropy loss for min-entropy $k \geq C \log^2 n$, some constant $C > 1$, as long as the security parameter $s$ satisfies $k \geq Cs^2$.

The next improvement in non-malleable extractors appeared in [CGL16], where Chattopadhyay, Goyal and Li constructed explicit non-malleable extractors with error $\varepsilon$, for min-entropy $k = \Omega(\log^2(n/\epsilon))$ and seed-length $d = O(\log^2(n/\epsilon))$. This gives an alternative protocol matching that of [Li15a]. Further improvements were obtained by Cohen [Coh16b, Coh16c], where he constructed non-malleable extractors with seed length $d = O(\log(n/\epsilon)\log((\log n)/\epsilon))$ and min-entropy $k = \Omega(\log(n/\epsilon)\log((\log n)/\epsilon))$; seed-length $O(\log n)$ and min-entropy $k = n/(\log n)^{O(1)}$; and seed length $d = O(\log n + \log^3(1/\epsilon))$ and min-entropy $k = \Omega(d)$. However, none of these improves the privacy amplification protocols in [Li15a].

Very recently, Chattopadhyay and Li [CL16] obtained an improved non-malleable extractor with error $\varepsilon$, for min-entropy $k = \log(n/\epsilon)2^{O(\sqrt{\log\log(n/\epsilon)})}$ and seed-length $d = \log(n/\epsilon)2^{O(\sqrt{\log\log(n/\epsilon)})}$, and min-entropy $k = O(\log n)$ and seed length $d = O(\log n)$ for error $\epsilon \geq 2^{-\log^{1-\beta} n}$ for any constant $0 < \beta < 1$. Independently, Cohen [Coh16a] also obtained a non-malleable extractor with error $\varepsilon$, for min-entropy $k = O(\log n) + \log(1/\epsilon)2^{O(\sqrt{\log\log(1/\epsilon)})}$ and seed-length $d = O(\log n) + \log(1/\epsilon)2^{O(\sqrt{\log\log(1/\epsilon)})}$. Both these constructions give two round privacy amplification protocols with optimal entropy loss, for security parameter $s$ up to $k/2^{O(\sqrt{\log k})}$.

**Seedless non-malleable extractors and non-malleable codes.** Seedless non-malleable extractors were first introduced by Cheraghchi and Guruswami [CG14b], in the context of non-malleable codes. Non-malleable codes, introduced by Dziembowski, Pietrzak and Wichs [DPW10], are a useful generalization of standard error correcting codes in the sense that they can handle a much larger class of attacks. Most notably, they can provide security guarantees even if the attacker can completely overwrite the codeword. Informally, a non-malleable code for a specific tampering family of tampering functions $\mathcal{F}$, consists of a randomized encoding function $E$ and a

deterministic decoding function $D$, such that if a codeword $E(x)$ is modified into $f(E(x))$ by some function $f \in \mathcal{F}$, then the decoded message $x' = D(f(E(x)))$ is either the original message $x$, or a completely unrelated message. The formal definition is given in Section 7. As shown in [DPW10], such non-malleable codes can be used in several applications in tamper-resilient cryptography.

While it can be seen that even non-malleable codes cannot exist if $\mathcal{F}$ is completely unrestricted, it is also known to exist for many broad tampering families. One of the most natural tampering families, and the most well studied, is the so called *split-state* model. Here, a $k$-bit message $x$ is encoded into $t$ parts of messages $y_1, \cdots, y_t$, each of length $n$. Now the adversary can arbitrarily tamper with each $y_i$ independently. In this case, the rate of the code is defined as $k/(tn)$.

This model arises in many applications naturally, for example when the different parts of messages $y_1, \cdots, y_t$ are stored in different parts of memory. It can also be viewed as a kind of "non-malleable secret sharing scheme". Clearly, the case of $t = 1$ corresponds to unrestricted tampering functions, and cannot be handled by non-malleable codes. Thus the case of $t = 2$ is the most useful and interesting setting. There has been a lot of work studying non-malleable codes in the $t$-split-state model. Since in this paper we focus on the information theoretic setting, we will only briefly review those previous work in the same setting.

The existence of non-malleable codes was first proved in [DPW10], and then Cheraghchi and Guruswami [CG14a] improved this result to show that the optimal rate of non-malleable codes in the 2-split-state model is 2. The first explicit construction appears in [DKO13], where the authors constructed explicit non-malleable codes for 1-bit messages in the split-state model. Subsequently, Aggarwal et. al [ADL14] constructed the first explicit non-malleable code for $k$-bit messages. Their encoding has message length $n = O(k^7 \log^7 k)$. This was later improved by Aggarwal [Agg14] to obtain $n = O(k^7)$.

Cheraghchi and Guruswami [CG14b] found a connection between non-malleable $t$-source extractors and non-malleable codes in the $t$-split state model. Their construction allows one to construct non-malleable codes in the $t$-split state model given sufficiently good non-malleable $t$-source extractors. However, they were not able to construct explicit non-malleable two-source extractors even for min-entropy $k = n$. Using this connection and techniques form additive combinatorics, Chattopadhyay and Zuckerman [CZ14] constructed a non-malleable 10-source extractor and a constant rate non-malleable code in the 10-split-state model. In a subsequent work, Chattopadhyay, Goyal and Li [CGL16] constructed the first explicit non-malleable two-source extractor for min-entropy $k = (1-\gamma)n$ with output $\Omega(k)$ and error $2^{-k^{\Omega(1)}}$, and used it to give an explicit non-malleable code in the 2-split state model with rate $n^{-\Omega(1)}$.

Finally, the work of Aggarwal et. al [ADKO15], has a construction which "achieves" a constant rate non-malleable code in the 2-split-state model. However, recently the author found an error in their proof (we briefly discuss the error in Appendix A), and thus this result does not hold. Currently, only non-malleable codes of rate $n^{-\Omega(1)}$ can be deduced from their work.

### 1.1 Our Results

We obtain improved results in all of the above problems. First, we have the following theorem which gives improved constructions of seeded non-malleable extractors.

**Theorem 1.7.** *There exists a constant $C > 1$ such that for any $n, k \in \mathbb{N}$ and $0 < \epsilon < 1$ with $k \geq C(\log n + \log\log(1/\epsilon)\log(1/\epsilon))$, there is an explicit strong seeded $(k, \epsilon)$ non-malleable extractor $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = C(\log n + \log\log(1/\epsilon)\log(1/\epsilon))$ and $m \geq k/4$.*

Combined with the protocol in [DW09], this gives the following theorem.

**Theorem 1.8.** *There exists a constant $0 < \alpha < 1$ such that for any $n, k \in \mathbb{N}$ and security parameter $s \leq \alpha k / \log k$, there is an explicit two-round privacy amplification protocol with entropy loss $O(\log n + s)$, in the presence of an active adversary.*

Combined with the techniques in [BADTS16], we obtain the following theorem which gives improved constructions of two-source extractors.

**Theorem 1.9.** *For every constant $\epsilon > 0$ there exists a constant $c > 1$ and an explicit two-source extractor $\mathsf{Ext} : (\{0,1\}^n)^2 \to \{0,1\}$ for min-entropy $k \geq c \log n \log\log n$, with error $\epsilon$.*

As a corollary, we obtain the following improved constructions of Ramsey graphs.

**Corollary 1.10.** *For every large enough integer $N$ there exists a (strongly) explicit construction of a $K$-Ramsey graph on $N$ vertices with $K = (\log N)^{O(\log\log\log N)}$*

Next we give an improved construction of a non-malleable two-source extractor.

**Theorem 1.11.** *There exists a constant $0 < \gamma < 1$ and a non-malleable two-source extractor for $(n, (1-\gamma)n)$ sources with error $2^{-\Omega(n/\log n)}$ and output length $\Omega(n)$.*

We give an algorithm to efficiently sample from the pre-image of this extractor, and together with the connection in [CG14b], we obtain the following theorem.

**Theorem 1.12.** *For any $n \in \mathbb{N}$ there exists an explicit non-malleable code with efficient encoder/decoder in the 2-split-state model with block length $2n$, rate $\Omega(1/\log n)$ and error $= 2^{-\Omega(n/\log n)}$.*

Finally, we use the non-malleable 10-source extractor in [CZ14] to obtain the following theorem.

**Theorem 1.13.** *For every constant $\epsilon > 0$ there exists a constant $c > 1$ and an explicit ten-source extractor $\mathsf{Ext} : (\{0,1\}^n)^{10} \to \{0,1\}$ for min-entropy $k \geq c \log n$, with error $\epsilon$.*

**Independent Work.** Independent of our work, and using different techniques, Cohen [Coh16d] obtained similar results for seeded non-malleable extractors and two-source extractors. Specifically, he constructed seeded non-malleable extractors for seed length and min-entropy $O(\log n) + \log(1/\epsilon)\text{polylog}\log(1/\epsilon)$, that outputs $\Omega(k)$ bits. He also constructed two-source extractors for min-entropy $\log n \text{polylog}\log n$.

## 1.2 Overview of The Constructions and Techniques

Here we give a brief overview of our constructions and the techniques. Both our constructions of seeded non-malleable extractor and seedless non-malleable extractor follow the high level framework of recent constructions [CGL16, Coh16b, Coh16c, CL16, Coh16a]. Specifically, we first obtain a small advice such that with high probability the untampered advice is different from the tampered version. The short size of the advice guarantees that even conditioned on the fixing of the advice, the seed and the source (or different sources) are still independent and have high min-entropy. We then use an improved correlation breaker with advice to obtain the output. Informally, given the advice, the correlation breaker does a series of computations using the inputs; and the output is

guaranteed to be close to uniform given the tampered output, if the advice is different from the tampered advice.

Take the seeded non-malleable extractor for example. It is well known that to achieve error $\epsilon$, one can use an advice of length $O(\log(n/\epsilon))$ (or even smaller, as shown in [Coh16a]), and length $\Omega(\log(1/\epsilon))$ is necessary. Moreover, this only costs $O(\log(n/\epsilon))$ bits in the seed and $O(\log(n/\epsilon))$ entropy in the source. We now turn to the part of the correlation breaker with advice. This part is going to follow the recent developments in [CL16, Coh16a], where (non-malleable) *independence preserving mergers* are used to construct the correlation breaker with advice. Specifically, let us briefly recall what is done in [CL16]. There, given the advice of length $L$, we first use an additional $O(\log(n/\epsilon))$ bits to create a matrix of $L$ rows, such that each row corresponds to a bit in the advice and each is uniform (but may be correlated with other rows). The property guaranteed is that on the bit that is different in the advice and the tampered advice, the corresponding row in the matrix is uniform even conditioned on the corresponding row in the tampered version of the matrix. Then, using the rest of the bits from the seed, we merge the matrix into one final row, while keeping this independence.

In [CL16], the construction first uses a basic merger, which uses $O(l \log(m/\epsilon))$ random bits to merge a matrix of $l$ rows, each row having length $m$. Then, one chooses a particular $l$ and applies the basic merger to the initial matrix of $L$ rows, merging $l$ rows each time. This takes $\log L/\log l$ steps. Each step one needs to use fresh random bits. However, since there is also a tampered seed, if each time we use the same number of fresh random bits, then they may already contain no entropy given the previously leaked tampered seeds. Therefore, in [CL16], each time the number of fresh random bits used is at least twice as large as the number of random bits used in the previous step. This means the number of random bis needed is going to grow exponentially, and eventually we need $2^{O(\log L/\log l)} l \log(m/\epsilon)$ random bits. A simple calculation shows that to minimize this quantity, we should choose $l$ such that $\log l = \sqrt{\log L}$ and this gives us $2^{O(\sqrt{\log L})} \log(m/\epsilon) = 2^{O(\sqrt{\log\log(n/\epsilon)})} \log(m/\epsilon)$ bits needed.

In this paper, we improve the merger in [CL16]. From the above discussion, one can see that if somehow we can get around the bottleneck of doubling the length of the random bits used each time, then ideally we would just need $O(l \log L/\log l \log(m/\epsilon))$ random bits. This quantity is minimized when $l$ is a constant (e.g., 2) and this gives us $O(\log L \log(m/\epsilon)) = O(\log\log(n/\epsilon))\log(m/\epsilon)$ random bits, which is much better than the previous one. How do we achieve this? Recall that previously the reason why we need to double the length of the random bits used each time, is that previously used bits from the tampered version can leak information about the current random bits of the untampered seed. If we can prevent this from happening, then we will be done. In other words, what we now need is to guarantee that each time the new random bits used in the seed is (close to) independent of the random bits previously used in the tampered version. Our crucial observation is that this is exactly a "look-ahead" property, and can be achieved by using alternating extraction.

This motivates the following construction. Let the source be $X$ and the seed be $Y$. After obtaining the advice, take a small slice $Y_1$ of $Y$ and use $Y_1$ to extract a small uniform output $Z$ from $X$. Use $Z$ and $Y$ (which still has a lot of entropy) to do an alternating extraction and output $\log L + 1$ random variables $R_i$. One can show that conditioned on the fixing of $Z$, these random variables are all deterministic functions of $Y$, and each $R_i$ is close to uniform conditioned on the previous ones and the previous tampered ones (i.e., they satisfy the look-ahead property). Now, we can use $R_1$ and $X$ (which, again, still has a lot of entropy) to create the initial matrix of $L$ rows, and then subsequently each time use a new $R_i$ to merge this matrix.

The above construction almost achieves what we want, except one problem. The problem is that the basic merger, which uses alternating extraction itself, only outputs say $0.2m$ bits if originally each row has $m$ bits (think of the non-malleable extractor case, which can output at most $k/2$ bits if the min-entropy is $k$). Thus, if we simply repeat the merging step for $\log L$ steps, then the length of the output will decrease to $2^{-O(\log L)}m$; and for this to be meaningful we would need $m \geq 2^{O(\log L)}$, which would make $m$ and also the min-entropy $k$ become at least $\text{poly}(L) = \text{polylog}(n/\epsilon)$. This is too large for our goal. Thus, we modify this construction so that we can compensate for the loss of output length each time. Specifically, after obtaining the advice, we first take a small slice $Y_1$ of $Y$ and use $Y_1$ to extract a small uniform output $Z$ from $X$. Note that conditioned on the fixing of $Y_1$, $Z$ is a deterministic function of $X$. Now we take a slightly larger slice $Y_2$ of $Y$, and a slice $Z_2$ of $Z$. Note that given $(Y_1, Y_2)$, $Y$ still has a lot of entropy. Similarly, given $(Y_1, Z_2)$, $Z$ still has a lot of entropy. We will now first use $Z_2$ and $Y$ to do an alternating extraction and output $2\log L + 1$ random variables $R_i$. We will also use $Y_2$ and $Z$ to do an alternating extraction and output $\log L + 1$ random variables $S_i$. One can show that conditioned on $(Y_1, Z)$, all the $R_i$ are deterministic functions of $Y$, and satisfy the look-ahead property. Similarly, conditioned on $(Y_1, Y_2)$, all the $S_i$ are deterministic functions of $X$, and satisfy the look-ahead property. We now use $S_0$ and $R_0$ (the first blocks in the sequences) to obtain the initial matrix, which conditioned on the fixing of $R_0$ is a deterministic function of $S_0$. Then, we repeat the merging for $\log L$ steps. Each step we will use two $R_i$'s and one $S_i$. Consider a particular step $i$. We first use $R_{2i-1}$ to merge the matrix, reducing the number of rows to a half. Note that conditioned on the fixing of $R_{2i-1}$, the output is a deterministic function of $S_{i-1}$. We then use each row of the output as a seed to extract from $R_{2i}$. Now conditioned on the previous matrix, the new output is a deterministic function of $R_{2i}$. Finally, we use each row of the new output as a seed to extract from $S_i$. Conditioned on the fixing of $R_{2i}$, the output becomes a deterministic function of $S_i$, and by choosing the length of each $S_i$ to be larger than $2m$ we can restore the length of each row in the matrix to $m$. This whole process still preserves the independence between the matrix and the tampered version of the matrix. We can thus repeat the process until we obtain the final output. Note that for all the alternating extraction, we can control the length of $Z$ and $S_i$, so that the number of random bits used is smaller than $O(\log(n/\epsilon))$. We also need to set $\epsilon$ to be slightly smaller than the error we want to achieve. Careful calculations show that we can achieve the seed length and entropy requirement in Theorem 1.7. By setting the parameters correctly, we can also ensure that the whole process described above does not consume much entropy, thus we can use the final output to extract from the original source and output $\Omega(k)$ bits.

The non-malleable two-source extractor follows essentially the same construction, except we now know that both sources already have min-entropy $(1-\gamma)n$. Thus, we can afford to set the error parameter to be $2^{-\Omega(n/\log n)}$.

**Efficient sampling.** The above non-malleable two-source extractor implies a non-malleable code in the 2-split-state model with rate $\Omega(1/\log n)$. However, to obtain an efficient encoder, we need to find a way to efficiently sample uniformly from the pre-image of any given output. Since the construction of the non-malleable two-source extractor is complicated and involves multi steps of alternating extraction etc., it appears that the sampling procedure may also be complicated. Indeed, in [CGL16] the sampling procedure consists of a series of carefully designed steps to "invert" each intermediate extraction step. Here, we show that in fact we can significantly simplify the sampling procedure. In fact, we are going to treat most of the details in the construction of the non-malleable

two-source extractor as a black box, and all we need are two ingredients from [CGL16]: First, a seeded extractor $\mathsf{IExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log(n/\epsilon))$ and $m = \Omega(d)$, such that for any fixed output $s$ and any fixed seed $r$, one can efficiently uniformly sample from the pre-image (this is because for any fixed seed, the output is a linear function of the input source), and the pre-image always has the same size. Second, to obtain the advice, first we take a small slice $X_1$ of the source $X$, and a small slice $Y_1$ of the source $Y$. Both slices have size $3\gamma n$ (assuming both sources have min-entropy $(1-\gamma)n$). We take the inner product of $X_1$ and $Y_1$, and use the output to sample $\Omega(n/\log n)$ coordinates from the Reed-Solomon encodings of both the rest part of $X$ and the rest part of $Y$. The advice $\alpha$ is obtained by concatenating $X_1$, $Y_1$ and the sampled coordinates. Now we slightly modify the non-malleable two-source extractor in the following way. We will take two other slices $Y_2$ and $Y_3$ of $Y$, with the guarantee that each has high min-entropy conditioned on previously leaked information, and the total length of $(Y_1, Y_2, Y_3)$ is less than $n/2$ (but still $\Omega(n)$). Similarly we take another slice $X_2$ of $X$, which has high min-entropy conditioned on previously leaked information, and the total length of $(X_1, X_2)$ is less than $n/2$ (but still $\Omega(n)$). Given the advice, we use $(X_2, Y_2)$ to run the non-malleable two source extractor described above, and obtain an output $V$. We then compute the final output $W = \mathsf{IExt}(Y_3, V)$. The non-malleable two-source extractor guarantees that $V$ is close to uniform given the tampered version, and this will be preserved in $W$.

Given any output $W$, we now briefly describe how to efficiently uniformly sample from the pre-image. We first uniformly generate $(X_1, Y_1, X_2, Y_2)$ and the advice $\alpha$. From these things we can compute the output $V$. Note that here we are treating the details in the construction of the non-malleable two-source extractor as a black box. Now, given $V$ and $W$, by the property of $\mathsf{IExt}$ we can efficiently sample $Y_3$, and the pre-image always has the same size. Finally, we need to sample the rest parts of $X$ and $Y$, given the variables we have obtained and $\alpha$. For this step, we note that once we have $(X_1, Y_1)$, we know the coordinates of the Reed-Solomon codes that we sampled, and these give us a system of linear equations. Note that we have at least $n/2$ free variables in both $X$ and $Y$, thus by setting the length of the advice appropriately (which is $\Omega(n)$) we can ensure that there are more variables in the system of equations than the number of equations. Therefore we can efficiently sample the pre-image by inverting the system of linear equations. Further note that the encoding matrix of the Reed-Solomon code has the property that regardless of the positions of the coordinates, as long as the number of sampled coordinates is the same, the encoding matrix always has the same rank. Thus the pre-image also has the same size regardless of the positions of the coordinates sampled. Therefore, altogether we can efficiently uniformly sample from the pre-image.

**Independent source extractor.** A corollary of the work of Ben-Aroya et. al [BADTS16] is that if one can construct seeded $t$-non-malleable extractor for some constant $t$ with error $\epsilon$, seed length and min-entropy $O(\log(n/\epsilon))$, then one also gets an explicit two-source extractor for min-entropy $O(\log n)$. The two-source extractor outputs one bit with any constant error. In this paper we show that we can reduce the task of constructing such seeded non-malleable extractor to the task of constructing non-malleable two-source extractors for $(n, (1-\gamma)n)$ sources with error $2^{-\Omega(n)}$, where $\gamma$ is any constant.

To see this, suppose we have such a non-malleable two-source extractor, then we can construct a seeded non-malleable extractor roughly as follows. Let the seed be $Y$ and the source be $X$. First, we can take a small slice of $Y$ and use it as a seed in an extractor, to convert $X$ into a close to

uniform string. Let the result be $\bar{X}$. Then, as usual, we obtain an advice $\alpha$ such that $\alpha \neq \alpha'$ with high probability, where $\alpha'$ is the tampered version of $\alpha$. Now, we take a small slice $Y_2$ of $Y$, and a small slice $X_2$ of $\bar{X}$, with the guarantee that both slices have entropy rate $> 1/2$. We take the inner product of $(X_2, Y_2)$, and use this output as an extractor to convert both $\bar{X}$ and $Y$ back into nearly uniform strings (the reason why we can do this is that the inner product is a two-source extractor strong in both sources). Let the outputs be $\tilde{X}$ and $\tilde{Y}$. We can now append $\alpha$ to both $\tilde{X}$ and $\tilde{Y}$. By setting the lengths appropriately we obtain two independent (conditioned on the fixing of previous random variables) $(m, (1-\gamma)m)$ sources, where $m = O(\log(n/\epsilon))$ as long as both $X$ and $Y$ have min-entropy at least $C \log(n/\epsilon)$ for some constant $C > 1$. We know that with high probability both sources will be different than their tampered version, thus we can now apply the non-malleable two-source extractor to get an output with error $\epsilon$.

The above construction is just for one tampering function, but we can use an argument similar to that used in [Li13a, Coh15] to gradually increase the resilience, until eventually the extractor works for $t$ tampering functions. This puts an $O(t^2)$ factor on the seed length and entropy requirement, which is still a constant if $t$ is a constant.

Clearly, the approach described above works not just for non-malleable extractors with optimal error, but works for any non-malleable extractor. Thus our non-malleable two-source extractor directly implies a two-source extractor for $(n, O(\log n \log\log n))$ sources. The approach also extends naturally to the case of non-malleable $(s+1)$-source extractor, which would give a seeded non-malleable extractor for $s$ independent sources. Thus, we can use the non-malleable 10-source extractor with optimal error in [CZ14], which gives a seeded non-malleable extractor for 9 independent sources. Together with the construction in [BADTS16] this gives an explicit extractor for 10 independent $(n, O(\log n))$ sources, which outputs one bit with any constant error.

**Organization.** The rest of the paper is organized as follows. We give some preliminaries in Section 2. We then define alternating extraction in Section 3, and non-malleable independence preserving merger in Section 4. In Section 5 we construct the new correlation breaker with advice. In Section 6 we present the seeded non-malleable extractor. In Section 7 we present non-malleable two-source extractors and non-malleable codes in the two-split-state model. Section 8 gives constructions of $t$-non-malleable extractors and applications to independent source extractors. Finally we conclude with some discussions and open problems in Section 9.

## 2 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.

### 2.1 Probability distributions

**Definition 2.1** (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 $h(x)$.

**Lemma 2.2.** *For any function $\alpha$ and two random variables $A, B$, we have $\Delta(\alpha(A), \alpha(B)) \leq \Delta(A, B)$.*

## 2.2 Somewhere Random Sources and Extractors

**Definition 2.3** (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 2.4.** (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 $\{0,1\}^d$,

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

## 2.3 Average conditional min-entropy

**Definition 2.5.** 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 2.6** ([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 2.7** ([DORS08])**.** *If a random variable $B$ has at most $2^\ell$ possible values, then $\widetilde{H}_\infty(A|B) \geq H_\infty(A) - \ell$.*

## 2.4 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 2.8.** *[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 2.9** ([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 $m \geq (1 - \alpha)k$. In addition, for any $\epsilon > 2^{-k/3}$ this gives a strong $(k, \epsilon)$ average case extractor with $m \geq k/2$.*

**Theorem 2.10** ([CG88])**.** *For every $0 < m < n$ there is an explicit two-source extractor $\mathsf{IP} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ based on the inner product function, such that if $X, Y$ are two independent $(n, k_1)$ and $(n, k_2)$ sources respectively, then*

$$(\mathsf{IP}(X,Y),X) \approx_\epsilon (U_m,X) \text{ and } (\mathsf{IP}(X,Y),Y) \approx_\epsilon (U_m,Y),$$

*where* $\epsilon = 2^{-\frac{k_1+k_2-n-m-1}{2}}$.

We need the following explicit construction of seedless non-malleable extractors in [CZ14].

**Theorem 2.11.** *There exists a constant* $\delta > 0$ *and an explicit* $(k,\epsilon)$*-seedless non-malleable extractor for* 10 *independent sources* $\mathsf{CZExt} : (\{0,1\}^n)^{10} \to \{0,1\}^m$ *with* $k = (1-\delta)n$, $\epsilon = 2^{-\Omega(n)}$ *and* $m = \Omega(k)$.

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

**Lemma 2.12** ([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 2.13.** *[Li13a] 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 2.14.** *[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* $|Z - U_n| \leq \epsilon^t$.

# 3 Alternating Extraction

An important ingredient in our construction is the following alternating extraction protocol, which was first introduced in [DP07], and then used a lot in constructions related to extractors (e.g., [DW09, Li13a]).

**Alternating Extraction.** Assume that we have two parties, Quentin and Wendy. Quentin has a source $Q$, Wendy has a source $W$. 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 $W$ is kept secret from Quentin. Let $\mathsf{Ext}_q$, $\mathsf{Ext}_w$ be strong seeded extractors with optimal parameters, such as that in Theorem 2.9. Let $s$ be an integer parameter for the protocol. For some integer parameter $\ell > 0$, the *alternating extraction protocol* is an interactive process between Quentin and Wendy that runs in $\ell$ steps.

In the first step, Quentin sends $S_1$ to Wendy, Wendy computes $R_1 = \mathsf{Ext}_w(W,S_1)$. 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 $s$ bits. In each subsequent step $i$, Quentin sends $S_i$ to Wendy, Wendy computes $R_i = \mathsf{Ext}_w(W,S_i)$. 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 $s$ bits. Therefore, this process produces the following sequence:

$$S_1, R_1 = \mathsf{Ext}_w(W,S_1), S_2 = \mathsf{Ext}_q(Q,R_1), \cdots, S_\ell = \mathsf{Ext}_q(Q,R_{\ell-1}), R_\ell = \mathsf{Ext}_w(W,S_\ell).$$

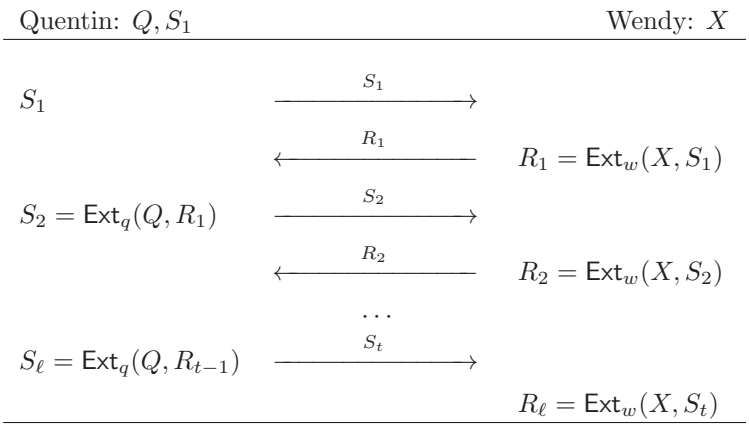


Figure 1: Alternating Extraction.

**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

$$\mathsf{laExt}(W, Y) = \mathsf{laExt}(W, (Q, S_1)) \stackrel{def}{=} R_1, \cdots, R_\ell.$$

The following lemma is a special case of Lemma 6.5 in [CGL16].

**Lemma 3.1.** *Let $W$ be an $(n_w, k_w)$-source and $W'$ be a random variable on $\{0,1\}^{n_w}$ that is arbitrarily correlated with $W$. Let $Y = (Q, S_1)$ such that $Q$ is a $(n_q, k_q)$-source, $S_1$ is a uniform string on $s$ bits, and $Y' = (Q', S_1')$ be a random variable arbitrarily correlated with $Y$, where $Q'$ and $S_1'$ are random variables on $n_q$ bits and $s$ bits respectively. Let $\mathsf{Ext}_q, \mathsf{Ext}_w$ be strong seeded extractors that extract $s$ bits from sources with min-entropy $k$ with error $\epsilon$ and seed length $s$. Suppose $(Y, Y')$ is independent of $(W, W')$, and $k_w, k_q \geqslant k + 2\ell s + 2\log(\frac{1}{\epsilon})$. Let $\mathsf{laExt}$ be the look-ahead extractor defined above using $\mathsf{Ext}_q, \mathsf{Ext}_w$, and $(R_1, \cdots, R_\ell) = \mathsf{laExt}(W, Y)$, $(R_1', \cdots, R_\ell') = \mathsf{laExt}(W', Y')$. Then for any $0 \leq j \leq \ell - 1$, we have*

$$(Y, Y', \{R_1, R_1', \cdots, R_j, R_j'\}, R_{j+1}) \approx_{\epsilon_1} (Y, Y', \{R_1, R_1', \cdots, R_j, R_j'\}, U_s),$$

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

# 4 Non-Malleable Independence Preserving Merger

We now describe the notion of *non-malleable independence preserving merger*, introduced in [CL16] based on the notion of independence preserving merger introduced in [CS16]. For simplicity we assume here we only have one adversary, which will be enough for our applications.

**Definition 4.1.** A $(L, d', \varepsilon)$-NIPM : $\{0,1\}^{Lm} \times \{0,1\}^d \to \{0,1\}^{m_1}$ satisfies the following property. Suppose

- $\mathbf{X}, \mathbf{X}'$ are random variables, each supported on boolean $L \times m$ matrices s.t for any $i \in [L]$, $\mathbf{X}_i = \mathbf{U}_m$,
- $\{\mathbf{Y}, \mathbf{Y}'\}$ is independent of $\{\mathbf{X}, \mathbf{X}'\}$, s.t $\mathbf{Y}, \mathbf{Y}'$ are each supported on $\{0,1\}^d$ and $H_\infty(\mathbf{Y}) \geqslant d'$,

- there exists an $h \in [L]$ such that $(\mathbf{X}_h, \mathbf{X}'_h) = (\mathbf{U}_m, \mathbf{X}'_h)$,

then

$$|(L, d', \varepsilon)\text{-NIPM}((\mathbf{X}, \mathbf{Y}), (L, d', \varepsilon)\text{-NIPM}(\mathbf{X}', \mathbf{Y}') - \mathbf{U}_{m_1}, (L, d', \varepsilon)\text{-NIPM}(\mathbf{X}', \mathbf{Y}')| \leqslant \epsilon.$$

We have the following construction and theorem.

***L*-Alternating Extraction** We extend the previous alternating extraction protocol by letting Quentin have access to $L$ sources $Q_1, \ldots, Q_L$ (instead of just $Q$) which have the same length. Now in the $i$'th round of the protocol, he uses $Q_i$ to produce the r.v $S_i = \mathsf{Ext}_q(Q_i, R_i)$. More formally, the following sequence of r.v's is generated: $S_1, R_1 = \mathsf{Ext}_w(W, S_1), S_2 = \mathsf{Ext}_q(Q_2, R_1), \ldots, R_{L-1} = \mathsf{Ext}_w(W, S_{\ell-1}), S_L = \mathsf{Ext}_q(Q_L, R_{L-1})$.

The NIPM is now constructed as follows. Let $S_1$ be a slice of $\mathbf{X}_1$ with length $O(\log(d/\varepsilon))$, then run the $L$-alternating extraction described above with $(Q_1, \ldots, Q_L) = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ and $W = \mathbf{Y}$. Finally output $S_L$.

**Theorem 4.2** ([CL16])**.** *There exists a constant $c > 0$ such that for all integers $m, d, d', L > 0$ and any $\epsilon > 0$, with $m \geqslant 4cL\log(d/\epsilon)$, $d' \geqslant 4cL\log(m/\epsilon)$, the above construction* $\text{NIPM} : (\{0,1\}^m)^\ell \times \{0,1\}^d \to \{0,1\}^{m_1}$ *has output length $m_1 \geq 0.2m$, such that if the following conditions hold:*

- $\mathbf{X}, \mathbf{X}'$ *are random variables, each supported on boolean $L \times m$ matrices s.t for any $i \in [L]$, $\mathbf{X}_i = \mathbf{U}_m$,*
- $\{\mathbf{Y}, \mathbf{Y}'\}$ *is independent of $\{\mathbf{X}, \mathbf{X}'\}$, s.t $\mathbf{Y}, \mathbf{Y}'$ are each supported on $\{0,1\}^d$ and $H_\infty(\mathbf{Y}) \geqslant d'$,*
- *there exists an $h \in [L]$ such that $(\mathbf{X}_h, \mathbf{X}'_h) = (\mathbf{U}_m, \mathbf{X}'_h)$,*

*then*

$$|\text{NIPM}((\mathbf{X}, \mathbf{Y}), \text{NIPM}((\mathbf{X}', \mathbf{Y}'), \mathbf{Y}, \mathbf{Y}' - \mathbf{U}_{m_1}, \text{NIPM}((\mathbf{X}', \mathbf{Y}'), \mathbf{Y}, \mathbf{Y}'| \leqslant L\epsilon.$$

# 5 Correlation Breaker with Advice

We now use the non-malleable independence preserving merger to construct an improved correlation breaker with advice. A correlation breaker, as its name suggests, uses independent randomness to break the correlations between several correlated random variables. A prototype correlation breaker was first constructed implicitly in the author's work [Li13a], and then later strengthened and formally defined in [Coh15]. A correlation breaker with advice additionally uses some string as an advice. This object was first introduced and used without its name in [CGL16], and then explicitly defined in [Coh16b]. We have the following definition.

**Definition 5.1** (Correlation breaker with advice)**.** A function

$$\mathsf{AdvCB} : \{0,1\}^n \times \{0,1\}^d \times \{0,1\}^a \to \{0,1\}^m$$

is called a $(k, \varepsilon)$-correlation breaker with advice if the following holds. Let $Y, Y'$ be $d$-bit random variables such that $Y$ is uniform. Let $X, X'$ be $n$-bit random variables with $H_\infty(X) \geq k$, such that $(X, X')$ is independent of $(Y, Y')$. Then, for any pair of distinct $a$-bit strings $\alpha, \alpha'$,

$$(\mathsf{AdvCB}(X,Y,\alpha),\mathsf{AdvCB}(X',Y',\alpha')) \approx_\varepsilon (U,\mathsf{AdvCB}(X',Y',\alpha')).$$

In addition, we say that $\mathsf{AdvCB}$ is strong if

$$(\mathsf{AdvCB}(X,Y,\alpha),\mathsf{AdvCB}(X',Y',\alpha'),Y,Y') \approx_\varepsilon (U,\mathsf{AdvCB}(X',Y',\alpha'),Y,Y').$$

For our construction we need the following flip-flop extraction scheme. The flip-flop function was constructed by Cohen [Coh15] using alternating extraction, based on a previous similar construction of the author [Li13a]. Subsequently, it was used in the construction of non-malleable extractors by Chattopadhyay, Goyal and Li [CGL16]. The flip-flop function is a basic version of correlation breaker, and (informally) uses an independent source $\mathbf{X}$ to break the correlation between two r.v's $\mathbf{Y}$ and $\mathbf{Y}'$, given an advice bit. We now describe this more formally.

**Theorem 5.2** ([Coh15, CGL16])**.** *There exists a constant $c_{5.2}$ such that for all $n > 0$ and any $\epsilon > 0$, there exists an explicit function* flip-flop $: \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, $m = 0.4k$, satisfying the following: Let $\mathbf{X}$ be an $(n,k)$-source, and $\mathbf{X}'$ be a random variable on $n$ bits arbitrarily correlated with $\mathbf{X}$. Let $\mathbf{Y}$ be an independent uniform seed on $d$ bits, and $\mathbf{Y}'$ be a random variable on $d$ bits arbitrarily correlated with $\mathbf{Y}$. Suppose $(\mathbf{X},\mathbf{X}')$ is independent of $(\mathbf{Y},\mathbf{Y}')$. If $k,d \geqslant C_{5.2}\log(n/\epsilon)$, then for any bit $b$,*

$$|\text{flip-flop}(\mathbf{X},\mathbf{Y},b),\mathbf{Y},\mathbf{Y}' - \mathbf{U}_m,\mathbf{Y},\mathbf{Y}'| \leqslant \epsilon.$$

*Furthermore, for any bits $b,b'$ with $b \neq b'$,*

$$|\text{flip-flop}(\mathbf{X},\mathbf{Y},b),\text{flip-flop}(\mathbf{X}',\mathbf{Y}',b'),\mathbf{Y},\mathbf{Y}' - \mathbf{U}_m,\text{flip-flop}(\mathbf{X}',\mathbf{Y}',b'),\mathbf{Y},\mathbf{Y}'| \leqslant \epsilon.$$

We construct a correlation breaker such that $X, X', Y, Y'$ are all on $d$ bits such that $H_\infty(X) \geq 0.9d$ and $H_\infty(Y) \geq 0.9d$. Using the above ingredients, our construction of the correlation breaker with advice is given below. For simplicity, when we say a strong seeded extractor for min-entropy $k$, we mean a strong average case seeded extractor for average conditional min-entropy $k$.

- Fix an error parameter $\epsilon'$ to be chosen later. Let $s$ be an integer such that $s \geq \max\{c\log(d/\epsilon'), 8c\log(3s/\epsilon')\}$ where $c$ is the maximum of the hidden constant in the seed length of the optimal seeded extractor in Theorem 2.9, and the two constants $c, c_{5.2}$ in Theorem 4.2 and Theorem 5.2.
- Let $\mathsf{Ext}$ be a strong seeded extractor which uses $r = c\log(3s/\epsilon')$ random bits to extract from an $(3s, 2c\log(3s/\epsilon'))$ source and outputs $r = c\log(3s/\epsilon')$ bits with error $\epsilon'$, from Theorem 2.9.
- Let $\mathsf{Ext}_w$, $\mathsf{Ext}_q$ be strong seeded extractors which use $s$ bits to extract from a $(d, 4s)$ source and outputs $3s$ bits with error $\epsilon'$.
- Let $\mathsf{Ext}'$ be a strong seeded extractor which uses $r = c\log(3s/\epsilon')$ random bits to extract from an $(3s, 1.5s)$ source and outputs $s$ bits with error $\epsilon'$, from Theorem 2.9.
- Let $\mathsf{Ext}''$ be a strong seeded extractor which uses $s \geq c\log(d/\epsilon')$ random bits to extract from a $(d, 0.15d)$ source and outputs $0.1d$ bits with error $\epsilon'$.
- Let $\mathsf{IP}$ be the two source extractor from Theorem 2.10, set up to extract from two $0.3d$-bit sources and output $0.05d$ bits.

1. Let $\ell = \log a$.[2] Let $X_1$ be a slice of $X$ with length $0.3d$, and $Y_1$ be a slice of $Y$ with length $0.3d$. Compute $Z = \mathsf{IP}(X_1, Y_1)$.

   Using $Z, Y$ as $Q, W$ (and $S_1$ is a small slice of $Q$) and $\mathsf{Ext}_w, \mathsf{Ext}_q$ as the extractors, run alternating extraction between $Z$ and $Y$ for $2\ell + 1$ steps, and output $R_0, R_1, R_2, \cdots, R_{2\ell} = \mathsf{laExt}(Y, Z)$, where each $R_i$ has $3s$ bits. Similarly, using $Z, X$ as $Q, W$ (and $S_1$ is a small slice of $Q$) and $\mathsf{Ext}_w, \mathsf{Ext}_q$ as the extractors, run alternating extraction between $Z$ and $X$ for $\ell + 1$ steps, and output $S_0, S_1, S_2, \cdots, S_\ell = \mathsf{laExt}(X, Z)$, where each $S_i$ has $3s$ bits.

2. Use $S_0, R_0, \alpha$ to obtain an $a \times s$ matrix $V^0$, where for any $i \in [a]$, $V_i^0 = \text{flip-flop}(S_0, R_0, \alpha_i)$ and outputs $s$ bits.

3. For $j = 1, \ldots, \ell$ do the following. Merge the matrix $V^{j-1}$ two rows by two rows: Note that $V^{j-1}$ has $a/2^{j-1}$ rows, for $i = 1, \ldots, a/2^j$, compute $\overline{V}_i^{j-1} = \text{NIPM}(V_{2i-1}^{j-1}, V_{2i}^{j-1}, R_{2j-1})$ which outputs $r$ bits, and $\tilde{V}_i^{j-1} = \mathsf{Ext}(R_{2j}, V_i'^{j-1})$ which has $r$ bits. Finally compute $V_i^j = \mathsf{Ext}'(S_j, \tilde{V}_i^{j-1})$ which has $s$ bits.

4. Compute $\hat{V} = \mathsf{Ext}''(X, \mathsf{Ext}_w(Y, V^\ell))$.

We now have the following lemma.

**Lemma 5.3.** *There exists a constant $C > 1$ such that for any $0 < \epsilon < 1/2$ and any $a, d \in \mathbb{N}$ such that $d \geq C \log a \log(da/\epsilon)$, there is an explicit construction of a function $\mathsf{AdvCB} : \{0,1\}^d \times \{0,1\}^d \times \{0,1\}^a \to \{0,1\}^{d/10}$ that satisfies the following. Let $Y, Y'$ be $d$-bit random variables such that $H_\infty(Y) \geq 0.9d$, and $X, X'$ be $d$-bit random variables with $H_\infty(X) \geq 0.9d$. Assume that $(X, X')$ is independent of $(Y, Y')$. Then, for any pair of distinct $a$-bit strings $\alpha, \alpha'$,*

$$(\mathsf{AdvCB}(X, Y, \alpha), \mathsf{AdvCB}(X', Y', \alpha'), Y, Y') \approx_\varepsilon (U, \mathsf{AdvCB}(X', Y', \alpha'), Y, Y').$$

*Proof.* We show that with appropriately chosen parameters $s, \epsilon'$ the above construction gives the desired correlation breaker with advice. We will use letters with prime to denote all the corresponding random variables produced by running the same algorithm on $(X', Y')$ instead of $(X, Y)$. Note that both $X_1$ and $Y_1$ has min-entropy at least $0.2d$. Thus by Theorem 2.10 we have that

$$(Z, X_1) - (U, X_1) \leq 2^{-\Omega(d)} \text{ and } (Z, Y_1) - (U, Y_1) \leq 2^{-\Omega(d)}.$$

We now fix $(Y_1, Y_1')$, and conditioned on this fixing $(Z, Z')$ is a deterministic function of $(X_1, X_1')$, thus independent of $(Y, Y')$. Moreover, $Z$ is close to uniform and the average conditional min-entropy of $Y$ is at least $0.9d - 2 \times 0.3d = 0.3d$.

Now by Lemma 3.1, as long as $0.3d \geq 4s + 2(2\ell+1)3s + 2\log(\frac{1}{\epsilon'})$ and $0.05d \geq 4s + 2(2\ell+1)3s + 2\log(\frac{1}{\epsilon'})$, we have that for any $0 \leq j \leq 2\ell - 1$,

$$(Z, Z', \{R_0, R_0', \cdots, R_j, R_j'\}, R_{j+1}) \approx_{O(\ell\epsilon')} (Z, Z', \{R_0, R_0', \cdots, R_j, R_j'\}, U_s).$$

By a hybrid argument and the triangle inequality, we have that

$$(Z, Z', R_0, R_0', \cdots, R_{2\ell}, R_{2\ell}') \approx_{O(\ell^2\epsilon')} (Z, Z', U_s, R_0', \cdots, U_s, R_{2\ell}'),$$

[2] Without loss of generality we assume that $a$ is a power of 2. Otherwise add 0 to the string until the length is a power of 2.

where each $U_s$ is independent of all the previous random variables (but may depend on later random variables). From now on, we will proceed as if each $R_{j+1}$ is uniform given $(Z, Z', \{R_0, R'_0, \cdots, R_j, R'_j\})$, since this only adds $O(\ell^2\epsilon')$ to the final error.

Note that conditioned on the fixing of $(Z, Z')$, we have that $\{(R_i, R'_i), i = 0, \ldots, 2\ell\}$ is a deterministic function of $(Y, Y')$, thus independent of $(X, X')$.

By symmetry, we can repeat the above argument while switching the role of $X$ and $Y$. Specifically, we can fix $(X_1, X'_1)$, and conditioned on this fixing $(Z, Z')$ is a deterministic function of $(Y_1, Y'_1)$, thus independent of $(X, X')$. Moreover, $Z$ is close to uniform and the average conditional min-entropy of $X$ is at least $0.9d - 2 \times 0.3d = 0.3d$.

Now again by Lemma 3.1, as long as $0.3d \geq 4s + 2(\ell+1)3s + 2\log(\frac{1}{\epsilon'})$ and $0.05d \geq 4s + 2(\ell+1)3s + 2\log(\frac{1}{\epsilon'})$, we have that for any $0 \leq j \leq \ell - 1$,

$$(Z, Z', \{S_0, S'_0, \cdots, S_j, S'_j\}, S_{j+1}) \approx_{O(\ell\epsilon')} (Z, Z', \{S_0, S'_0, \cdots, S_j, S'_j\}, U_s).$$

By a hybrid argument and the triangle inequality, we have that

$$(Z, Z', S_0, S'_0, \cdots, S_\ell, S'_\ell) \approx_{O(\ell^2\epsilon')} (Z, Z', U_s, S'_0, \cdots, U_s, S'_\ell),$$

where each $U_s$ is independent of all the previous random variables (but may depend on later random variables). From now on, we will proceed as if each $S_{j+1}$ is uniform given $(Z, Z', \{S_0, S'_0, \cdots, S_j, S'_j\})$, since this only adds $O(\ell^2\epsilon')$ to the final error.

Note that now conditioned on the fixing of $(Z, Z')$, we have that $\{(S_i, S'_i), i = 0, \ldots, \ell\}$ is a deterministic function of $(X, X')$, thus independent of $(Y, Y')$. Therefore, we can conclude that conditioned on the fixing of $(X_1, X'_1, Y_1, Y'_1, Z, Z')$, we have that $\{(R_i, R'_i), i = 0, \ldots, 2\ell\}$ is a deterministic function of $(Y, Y')$, and $\{(S_i, S'_i), i = 0, \ldots, \ell\}$ is a deterministic function of $(X, X')$, thus they are independent. Moreover each $R_i$ and $S_i$ is close to uniform given the previous random variables.

We now have the following claim.

**Claim 5.4.** *For all $i \in [a]$ we have that*

$$\left|V_i^0 - U_s\right| \leqslant \epsilon'.$$

*Furthermore, there exists an $i \in [a]$ such that*

$$\left|(V_i^0, V_i'^0, R_0, R'_0) - (U_s, V_i'^0, R_0, R'_0)\right| \leqslant \epsilon'.$$

Indeed, since $\alpha \neq \alpha'$ there exists an $i \in [a]$ such that $\alpha_i \neq \alpha'_i$. Thus by Theorem 5.2, and noticing that $3s \geq C_{5.2}\log(3s/\epsilon')$, the claim follows. Furthermore, notice that now conditioned on the fixing of $(R_0, R'_0)$, $(V^0, V'^0)$ is a deterministic function of $(S_0, S'_0)$, and thus independent of $\{(R_i, R'_i), i = 1, \ldots, 2\ell\}$. We now have the following claim.

**Claim 5.5.** *Assume that for some $j \leq \ell$, we have that for all $i$,*

$$\left|(V_i^j, \{R_0, R'_0, \cdots, R_{2j}, R'_{2j}\}) - (U_s, \{R_0, R'_0, \cdots, R_{2j}, R'_{2j}\})\right| \leqslant \epsilon_j.$$

*Furthermore there exists an $i$ such that*

$$\left|(V_i^j, V_i'^j, \{R_0, R'_0, \cdots, R_{2j}, R'_{2j}\}) - (U_s, V_i'^j, \{R_0, R'_0, \cdots, R_{2j}, R'_{2j}\})\right| \leqslant \epsilon_j.$$

*Then for all $i$, we have that*

$$\left|(V_i^{j+1}, \{R_0, R'_0, \cdots, R_{2(j+1)}, R'_{2(j+1)}\}) - (U_s, \{R_0, R'_0, \cdots, R_{2(j+1)}, R'_{2(j+1)}\})\right| \leqslant 2(\epsilon_j + 2\epsilon').$$

*Furthermore there exits an $i$ such that*

$$\left|(V_i^{j+1}, V_i'^{j+1}, \{R_0, R'_0, \cdots, R_{2(j+1)}, R'_{2(j+1)}\}) - (U_s, V_i'^{j+1}, \{R_0, R'_0, \cdots, R_{2(j+1)}, R'_{2(j+1)}\})\right| \leqslant 2(\epsilon_j + 2\epsilon').$$

To see the claim, we focus on the index $i$ where the corresponding row $V_i^j$ is close to uniform given $V_i'^j$. The properties of the other rows can be obtained using similar and simpler arguments. Notice that conditioned on the fixing of $\{R_0, R'_0, \cdots, R_{2j}, R'_{2j}\}$, we have that $(V^j, V'^j)$ is a deterministic function of $(S_0, S'_0, \cdots, S_j, S'_j)$, and thus independent of $(R_{2j+1}, R'_{2j+1})$. Furthermore, by the property of the look-ahead extractor, we know that $R_{2j+1}$ is uniform. Now by Theorem 4.2, and noticing that $s \geq 8c\log(3s/\epsilon')$, we know that whenever the NIPM merges the two rows in which one row of $V^j$ is uniform given the corresponding row of $V'^j$, the output obtained from $V^j$ will be uniform given the output obtained from $V'^j$. Thus, there exists an $i$ such that

$$\left|(\overline{V}_i^j, \overline{V'}_i^j, R_{2j+1}, R'_{2j+1}) - (U_r, \overline{V'}_i^j, R_{2j+1}, R'_{2j+1})\right| \leqslant 2\epsilon_j + 2\epsilon'.$$

Now we fix $(R_{2j+1}, R'_{2j+1})$, and conditioned on this fixing $(\overline{V}^j, \overline{V'}^j)$ is a deterministic function of $(S_0, S'_0, \cdots, S_j, S'_j)$, and thus independent of $(R_{2(j+1)}, R'_{2(j+1)})$. Moreover now again by the property of the look-ahead extractor, we know that $R_{2(j+1)}$ is uniform. Therefore, we can first fix $\overline{V'}_i^j$ and then $\tilde{V'}_i^j = \mathsf{Ext}(R'_{2(j+1)}, V_i'^j)$. Conditioned on this fixing we have that $\overline{V}_i^j$ is still uniform, and that $R_{2(j+1)}$ has average conditional min-entropy at least $3s - r = 3s - c\log(3s/\epsilon') \geq 23c\log(2s/\epsilon')$. Therefore, by Theorem 2.9 we have that

$$\left|(\tilde{V}_i^j, \tilde{V'}_i^j, \overline{V}_i^j, \overline{V'}_i^j) - (U_r, \tilde{V'}_i^j, \overline{V}_i^j, \overline{V'}_i^j)\right| \leq \epsilon'.$$

Now we can fix $(\overline{V}_i^j, \overline{V'}_i^j)$ and conditioned on this fixing, $(\tilde{V}_i^j, \tilde{V'}_i^j)$ is a deterministic function of $(R_{2(j+1)}, R'_{2(j+1)})$, and thus independent of $(S_{j+1}, S'_{j+1})$. Thus we can first fix $\tilde{V'}_i^j$ and then $V_i'^{j+1} = \mathsf{Ext}'(S'_{j+1}, \tilde{V'}_i^j)$. Note that after this fixing $\tilde{V}_i^j$ is still close to uniform, moreover the average conditional min-entropy of $S_{j+1}$ is at least $3s - s = 2s$. Thus by Theorem 2.9 we have that

$$\left|(V_i^{j+1}, V_i'^{j+1}, \tilde{V}_i^j, \tilde{V'}_i^j) - (U_s, V_i'^{j+1}, \tilde{V}_i^j, \tilde{V'}_i^j)\right| \leq \epsilon'.$$

Note that conditioned on the fixing of $(\tilde{V}_i^j, \tilde{V'}_i^j)$, we have that $(V_i^{j+1}, V_i'^{j+1})$ is a deterministic function of $(S_{j+1}, S'_{j+1})$, and thus independent of $(R_{2(j+1)}, R'_{2(j+1)})$. Since we have fixed all the $\{R_0, R'_0, \cdots, R_{2j}, R'_{2j}\}$ before, by adding all the errors we obtain that

$$\left|(V_i^{j+1}, V_i'^{j+1}, \{R_0, R'_0, \cdots, R_{2(j+1)}, R'_{2(j+1)}\}) - (U_s, V_i'^{j+1}, \{R_0, R'_0, \cdots, R_{2(j+1)}, R'_{2(j+1)}\})\right| \leqslant 2(\epsilon_j + 2\epsilon').$$

Now note that by the end of the iteration of step 3, $V^\ell$ has only one row. From Claim 5.5 we see that (by solving the recursion of the errors)

$$\left|(V^\ell, V'^\ell, \{R_0, R'_0, \cdots, R_{2\ell}, R'_{2\ell}\}) - (U_s, V'^\ell, \{R_0, R'_0, \cdots, R_{2\ell}, R'_{2\ell}\})\right| \leqslant 10a\epsilon'.$$

Note that conditioned on the fixing of $X_1, Y_1, X'_1, Y'_1, \{R_0, R'_0, \cdots, R_{2\ell}, R'_{2\ell}\}$, we have that $(V^\ell, V'^\ell)$ is a deterministic function of $(X, X')$, and thus independent of $(Y, Y')$. Furthermore the average conditional min-entropy of $Y$ is at least $0.9d - 2 \times 0.3d - 2(2\ell+1)3s = 0.3d - (12\ell+6)s$. Thus we can first fix $V'^\ell$ and then $\mathsf{Ext}_w(Y, V'^\ell)$, and conditioned on this fixing we have that $V^\ell$ is still close to uniform and independent of $Y$, and the average conditional min-entropy of $Y$ is at least $0.3d - (12\ell + 9)s$. Now as long as $0.3d - (12\ell + 9)s \geq 4s$, by Theorem 2.9 we have that

$$\left|\mathsf{Ext}_w(Y, V^\ell), \mathsf{Ext}_w(Y', V'^\ell), V^\ell, V'^\ell) - (U_{3s}, \mathsf{Ext}_w(Y', V'^\ell), V^\ell, V'^\ell\})\right| \leqslant \epsilon'.$$

Finally, notice that conditioned on the further fixing of $V^\ell, V'^\ell$, we have that $(\mathsf{Ext}_w(Y, V^\ell), \mathsf{Ext}_w(Y, V'^\ell))$ is a deterministic function of $(Y, Y')$, and thus independent of $(X, X')$. Furthermore the average conditional min-entropy of $X$ is at least $0.9d - 2 \times 0.3d - 2(\ell + 1)3s = 0.3d - (6\ell + 6)s$. Thus we can first fix $\mathsf{Ext}_w(Y', V'^\ell)l$ and then $\hat{V}' = \mathsf{Ext}''(X', \mathsf{Ext}_w(Y, V'^\ell))$, and conditioned on this fixing we have that $\mathsf{Ext}_w(Y, V^\ell)$ is still close to uniform and independent of $X$, and the average conditional min-entropy of $X$ is at least $0.3d - (6\ell + 6)s - 0.1d = 0.2d - (6\ell + 6)s$. Thus as long as $0.2d - (6\ell + 6)s \geq 0.15d$, Theorem 2.9 we have that

$$\left|\hat{V}, \hat{V}', \mathsf{Ext}_w(Y, V^\ell), \mathsf{Ext}_w(Y', V'^\ell)) - (U_{0.1d}, \hat{V}', \mathsf{Ext}_w(Y, V^\ell), \mathsf{Ext}_w(Y', V'^\ell))\right| \leqslant \epsilon'.$$

Note that now conditioned on the fixing of $(\mathsf{Ext}_w(Y, V^\ell), \mathsf{Ext}_w(Y', V'^\ell))$, we have that $(\hat{V}, \hat{V}')$ is a deterministic function of $(X, X')$, and thus independent of $(Y, Y')$. Therefore by adding back all the errors we obtain

$$\left|\hat{V}, \hat{V}', Y, Y') - (U_{0.1d}, \hat{V}', Y, Y')\right| \leqslant \epsilon_1,$$

where $\epsilon_1 = (10a + 2)\epsilon' + O(\ell^2\epsilon') + 2^{-\Omega(d)}$.

Next, in order for all the entropy requirement to hold, we need the following conditions.

$$s \geq \max\{c\log(d/\epsilon'), 8c\log(3s/\epsilon')\}, \text{ and } 0.05d \geq 4s + 2(2\ell + 1)3s + 2\log(\frac{1}{\epsilon'})$$

$$0.3d - (12\ell + 9)s \geq 4s, \text{ and } 0.2d - (6\ell + 6)s \geq 0.15d.$$

The above conditions are satisfied if the following conditions are satisfied.

$$d \geq 240(\ell + 1)s, \text{ and } s \geq 8c\log(d/\epsilon').$$

Under this condition, we see that $2^{-\Omega(d)} \leq \epsilon'$, and since $\ell = \log a$ we have that $\ell^2 = O(a)$. Thus the total error is $\epsilon_1 = O(a)\epsilon'$. Therefore, to make $\epsilon_1 = \epsilon$, we can set $\epsilon' = \epsilon/(c'a)$ for some constant $c' > 0$. We can now set $s = 9c\log(d/\epsilon') = 9c\log(c'da/\epsilon)$, and the conditions are satisfied as long as $d \geq C\ell\log(da/\epsilon) = C\log a\log(da/\epsilon)$ for some constant $C > 1$.

□

## 6 The Seeded Non-Malleable Extractor

In this section we construct our improved seeded non-malleable extractor. First we need the following advice generator from [CGL16]

**Theorem 6.1** ([CGL16]). *There exist a constant $c > 0$ such that for all $n > 0$ and any $\epsilon > 0$, there exists an explicit function $\mathsf{AdvGen} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^L$ with $L = c\log(n/\epsilon)$ satisfying the following: Let $X$ be an $(n,k)$-source, and $Y$ be an independent uniform seed on $d$ bits. Let $Y'$ be a random variable on $d$ bits s.t $Y' \neq Y$, and $(Y,Y')$ is independent of $X$. Then with probability at least $1-\epsilon$, $\mathsf{AdvGen}(X,Y) \neq \mathsf{AdvGen}(X,Y')$. Moreover, there is a deterministic function $g$ such that $\mathsf{AdvGen}(X,Y)$ is computed as follows. Let $Y_1$ be a small slice of $Y$ with length $O(\log(n/\epsilon))$, compute $Z_1 = \mathsf{Ext}(X,Y_1)$ where $\mathsf{Ext}$ is an optimal seeded extractor from Theorem 2.9 which outputs $O(\log(n/\epsilon))$ bits. Finally compute $Y_2 = g(Y,Z_1)$ which outputs $O(\log(1/\epsilon))$ bits and let $\mathsf{AdvGen}(X,Y) = (Y_1,Y_2)$.*

The construction of the non-malleable extractor is as follows.

- Let $\epsilon' = \epsilon/10$. Assume $k \geq 6d$.
- Let $\mathsf{Ext}$ be a strong seeded extractor from Theorem 2.9, which uses $O(\log(n/\epsilon'))$ bits to extract from an $(n,k/3)$ source and outputs $k/4$ bits with error $\epsilon'$.
- Let $\mathsf{Ext}'$ be a strong seeded extractor from Theorem 2.9, which uses $O(\log(n/\epsilon'))$ bits to extract from an $(n,k)$ source and outputs $d$ bits with error $\epsilon'$.
- Let $\mathsf{AdvGen}$ be the advice generator from Theorem 6.1, with error $\epsilon'$.
- Let $\mathsf{AdvCB}$ be the correlation breaker with advice from Lemma 5.3, with error $\epsilon'$.

1. Compute $\mathsf{AdvGen}(X,Y)$ with error $\epsilon'$. Specifically, first compute $X_1 = \mathsf{Ext}'(X,Y_1)$, except now it outputs $Z = \mathsf{Ext}(X,Y_1)$ with $d$ bits. Let $Z_1$ be a slice of $Z$ with $O(\log(n/\epsilon'))$ bits and as in Theorem 6.1, compute $Y_2 = g(Y,Z_1)$ which outputs $O(\log(1/\epsilon'))$ bits and let $\mathsf{AdvGen}(X,Y) = (Y_1,Y_2) = \alpha$.
2. Compute $V = \mathsf{AdvCB}(Y,Z,\alpha)$ which outputs $d/10$ bits.
3. Output $W = \mathsf{Ext}(X,V)$ which outputs $k/4$ bits

We now have the following theorem.

**Theorem 6.2.** *There exists a constant $C > 1$ such that for any $n,k \in \mathbb{N}$ and $0 < \epsilon < 1$ with $k \geq C(\log n + \log\log(1/\epsilon)\log(1/\epsilon))$, there is an explicit construction of a strong seeded $(k,\epsilon)$ non-malleable extractor $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = C(\log n + \log\log(1/\epsilon)\log(1/\epsilon))$ and $m \geq k/4$.*

*Proof.* Again, we use letters with prime to denote random variables produced with $(X,Y')$ instead of $(X,Y)$. First note that by Theorem 2.9, we have that

$$(Z,Y_1) \approx_{\epsilon'} (U_d,Y_1).$$

We will now proceed as if $Z$ is uniform given $Y_1$, since this only adds error $\epsilon'$. We now fix $(Y_1, Y_1')$. Note that conditioned on this fixing, $(Z, Z')$ is a deterministic function of $X$, and thus independent of $(Y, Y')$. Moreover by Lemma 2.12 with probability $1-\epsilon'$, the min-entropy of $Y$ is at least $d - O(\log(n/\epsilon'))$. Now we fix $(Z_1, Z_1')$, and note that conditioned on this fixing, $(Y_2, Y_2')$ is a deterministic function of $(Y, Y')$, and thus independent of $(X, Z, Z')$. Moreover again by by Lemma 2.12 with probability $1-\epsilon'$, the min-entropy of $Z$ is at least $d - O(\log(n/\epsilon'))$. Finally we fix $(Y_2, Y_2')$. Note that conditioned on this fixing, $(Y, Y')$ is still independent of $(X, Z, Z')$. Moreover by Lemma 2.12 with probability $1-\epsilon'$, the min-entropy of $Y$ is at least $d - O(\log(n/\epsilon'))$. Also note that by Theorem 6.1, with probability at least $1-\epsilon'$ over the fixing of $(Y_1, Z_1, Y_2, Y_1', Z_1', Y_2')$, we have that $\alpha = (Y_1, Y_2) \neq (Y_1', Y_2') = \alpha'$. Thus, as long as $d \geq C \log(n/\epsilon')$ for some constant $C > 1$, altogether we can conclude that with probability at least $1 - 4\epsilon'$, we have that

- $\alpha \neq \alpha'$, where $\alpha, \alpha'$ each has $a = c \log(n/\epsilon')$ bits.
- $X$ is still independent of $(Y, Y')$, and $(Z, Z')$ is a deterministic function of $X$.
- $H_\infty(Y) \geq 0.9d$ and $H_\infty(Z) \geq 0.9d$.

Thus, as long as $d \geq C' \log a \log(da/\epsilon')$ where $C'$ is the constant in Lemma 5.3, we have that

$$(V, V', Z, Z') \approx_{\epsilon'} (U, V', Z, Z').$$

Note that conditioned on the fixing of $(Z, Z')$, we have that $(V, V')$ is a deterministic function of $(Y, Y')$, and thus independent of $X$. Moreover the average conditional min-entropy of $X$ is at least $k - 2d \geq 2k/3$. Thus now we can first fix $V'$ and then $W' = \mathsf{Ext}(X, V')$. Note that after this fixing $X$ and $(Y, Y', V)$ are still independent. Moreover $V$ is still close to uniform and the average conditional min-entropy of $X$ is at least $2k/3 - k/4 > k/3$. Thus by Theorem 2.9 we have that

$$(W, W', V, V') \approx_{\epsilon'} (U, W', V, V').$$

Note that conditioned on the fixing of $(V, V')$, we have that $(W, W')$ is a deterministic function of $X$, thus independent of $(Y, Y')$. Therefore by adding back all the errors we get that

$$(W, W', Y, Y') \approx_{7\epsilon'} (U, W', Y, Y').$$

Since $\epsilon' = \epsilon/10$ we have that

$$(W, W', Y, Y') \approx_{\epsilon} (U, W', Y, Y').$$

Now let's decide the seed length $d$. We need to have that

$$d \geq C \log(n/\epsilon') \text{ and } d \geq C' \log a \log(da/\epsilon'),$$

where $a = c \log(n/\epsilon')$ and $\epsilon' = \epsilon/10$. Since our new non-malleable extractor is better than the construction in [CGL16], which has seed length $d = O(\log^2(n/\epsilon'))$, we can first assume that $d = O(\log^2(n/\epsilon))$ and we will use the inequality $d \geq C' \log a \log(da/\epsilon')$ to compute the minimum $d$ and verify the condition that $d = O(\log^2(n/\epsilon'))$ does hold.

In this case, we see that $\log(da/\epsilon') = \log(O(\log^3(n/\epsilon')/\epsilon')) = O(\log\log(n/\epsilon') + \log(1/\epsilon'))$, and $\log a = O(\log\log(n/\epsilon'))$. Thus we need

$$d \geq C_1((\log\log(n/\epsilon'))^2 + \log\log(n/\epsilon')\log(1/\epsilon'))$$

for some constant $C_1 > 1$.

Note that if $\epsilon' < 1/n$, then $\log\log(n/\epsilon') = O(\log\log(1/\epsilon'))$ and thus $(\log\log(n/\epsilon'))^2 < \log\log(n/\epsilon')\log(1/\epsilon')$; and if $\epsilon' \geq 1/n$ then $(\log\log(n/\epsilon'))^2 < \log n$. Thus we have

$$(\log\log(n/\epsilon'))^2 < \log n + \log\log(n/\epsilon')\log(1/\epsilon').$$

Now consider $\log\log(n/\epsilon')\log(1/\epsilon')$. We have that

$$\log\log(n/\epsilon')\log(1/\epsilon') \leq \log(\log n \log(1/\epsilon'))\log(1/\epsilon') = (\log\log n + \log\log(1/\epsilon'))\log(1/\epsilon').$$

Now if $\epsilon' < 2^{-\log n/\log\log n}$, then we have that $\log\log(1/\epsilon') > \log\log n - \log\log\log n > 0.5\log\log n$. Thus in this case we have that

$$\log\log(n/\epsilon')\log(1/\epsilon') \leq 3\log\log(1/\epsilon')\log(1/\epsilon').$$

On the other hand, if $\epsilon' \geq 2^{-\log n/\log\log n}$, then we have that

$$\log\log n \log(1/\epsilon') \leq \log n.$$

Thus combining the two cases we have that

$$\log\log(n/\epsilon')\log(1/\epsilon') \leq \log n + 3\log\log(1/\epsilon')\log(1/\epsilon').$$

Altogether we have

$$(\log\log(n/\epsilon'))^2 + \log\log(n/\epsilon')\log(1/\epsilon') \leq 3\log n + 6\log\log(1/\epsilon')\log(1/\epsilon').$$

Thus, it suffices to set

$$d = O(\log n + \log\log(1/\epsilon')\log(1/\epsilon')) = O(\log n + \log\log(1/\epsilon)\log(1/\epsilon)).$$

□

# 7 Non-Malleable Two-Source Extractor and Non-Malleable Code

Formally, non-malleable codes are defined as follows.

**Definition 7.1.** [ADKO15] Let $\mathsf{NM}_k$ denote the set of trivial manipulation functions on $k$-bit strings, which consists of the identity function $I(x) = x$ and all constant functions $f_c(x) = c$, where $c \in \{0,1\}^k$. Let $E : \{0,1\}^k \to \{0,1\}^m$ be an efficient randomized *encoding* function, and $D : \{0,1\}^m \to \{0,1\}^k$ be an efficient deterministic *decoding* function. Let $\mathcal{F} : \{0,1\}^m \to \{0,1\}^m$ be some class of functions. We say that the pair $(E, D)$ defines an $(\mathcal{F}, k, \epsilon)$-*non-malleable code*, if for all $f \in \mathcal{F}$ there exists a probability distribution $G$ over $\mathsf{NM}_k$, such that for all $x \in \{0,1\}^k$, we have

$$|D(f(E(x))) - G(x)| \leq \epsilon.$$

**Remark 7.2.** The above definition is slightly different form the original definition in [DPW10]. However, [ADKO15] shows that the two definitions are equivalent.

We will mainly be focusing on the following family of tampering functions in this paper.

**Definition 7.3.** Given any $\ell > 1$, let $\mathcal{S}_n^\ell$ denote the tampering family in the *$\ell$-split-state-model*, where the adversary applies $\ell$ arbitrarily correlated functions $h_1, \cdots, h_\ell$ to $\ell$ separate, $n$-bit parts of string. Each $h_i$ can only be applied to the $i$-th part individually.

Note that although the functions $h_1, \cdots, h_\ell$ can be correlated, their correlation does not depend on the original codewords. Thus, they are a convex combination of independent functions, applied to each part of the codeword. Thus, without loss of generality, hereafter we may assume that each $h_i$ is an independent function acting on the $i$-th part of the codeword individually. In this paper we will mainly consider the case of $\ell = 2$, i.e., the two-split-state model.

The following theorem was proved by Cheraghchi and Gursuswami [CG14b], which establishes a connection between seedless non-malleable extractors and non-malleable codes.

**Theorem 7.4.** *Let* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *be a polynomial time computable seedless 2-non-malleable extractor at min-entropy $n$ with error $\epsilon$. Then there exists an explicit non-malleable code with an efficient decoder in the 2-split-state model with block length $= 2n$, rate $= \frac{m}{2n}$ and error $= 2^{m+1}\epsilon$.*

Using the non-malleable extractor, the non-malleable code in the 2-split-state model is constructed as follows: For any message $s \in \{0,1\}^m$, the encoder $\mathsf{Enc}(s)$ outputs a uniformly random string from the set $\mathsf{nmExt}^{-1}(s) \subset \{0,1\}^{2n}$. For any codeword $c \in \{0,1\}^{2n}$, the decoder $\mathsf{Dec}$ outputs $\mathsf{nmExt}(c)$. Thus, for the encoder to be efficient we need to be able to efficiently uniformly sample from the pre-image of any output of the extractor. We will now first describe our construction of the non-malleable extractor and then show how to efficiently uniformly sample from the pre-image.

## 7.1 The construction and the analysis of the extractor

We have the following construction of a non-malleable two-source extractor for two $(n, (1-\gamma)n)$ sources, where $0 < \gamma < 1$ is some constant. First we need the following construction of an "invertible" linear seeded extractor.

**Theorem 7.5.** *There exists a constant $0 < \alpha < 1$ such that for any $n \in \mathbb{N}$ and $2^{-\alpha n} < \epsilon < 1$ there exists a linear seeded strong extractor $\mathsf{IExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{0.3d}$ with $d = O(\log(n/\epsilon))$ and the following property. If $X$ is a $(n, 0.9n)$ source and $R$ is an independent uniform seed on $\{0,1\}^d$, then*

$$|(\mathsf{IExt}(X,R),R) - (U_{0.3d},R)| \leq \epsilon.$$

*Furthermore for any $s \in \{0,1\}^{0.3d}$ and any $r \in \{0,1\}^d$, $|\mathsf{IExt}(\cdot,r)^{-1}(s)| = 2^{n-0.3d}$.*

To prove the theorem we need the following definitions and theorems.

**Definition 7.6** (Averaging sampler [Vad04])**.** A function $\mathsf{Samp} : \{0,1\}^r \to [n]^t$ is a $(\mu,\theta,\gamma)$ averaging sampler if for every function $f : [n] \to [0,1]$ with average value $\frac{1}{n}\sum_i f(i) \geqslant \mu$, it holds that

$$\Pr_{i_1,\ldots,i_t \leftarrow \mathsf{Samp}(U_R)} \left[ \frac{1}{t} \sum_i f(i) < \mu - \theta \right] \leq \gamma.$$

$\mathsf{Samp}$ has distinct samples if for every $x \in \{0,1\}^r$, the samples produced by $\mathsf{Samp}(x)$ are all distinct.

**Theorem 7.7** ([Vad04]). *Let $1 \geq \delta \geq 3\tau > 0$. Suppose that $\mathsf{Samp} : \{0,1\}^r \to [n]^t$ is an $(\mu, \theta, \gamma)$ averaging sampler with distinct samples for $\mu = (\delta - 2\tau)/\log(1/\tau)$ and $\theta = \tau/\log(1/\tau)$. Then for every $\delta n$-source $X$ on $\{0,1\}^n$, the random variable $(U_r, X_{Samp(U_r)})$ is $(\gamma + 2^{-\Omega(\tau n)})$-close to $(U_r, W)$ where for every $a \in \{0,1\}^r$, the random variable $W|_{U_r=a}$ is $(\delta - 3\tau)t$-source.*

**Theorem 7.8** ([Vad04]). *For every $0 < \theta < \mu < 1$, $\gamma > 0$, and $n \in \mathbb{N}$, there is an explicit $(\mu, \theta, \gamma)$ averaging sampler $\mathsf{Samp} : \{0,1\}^r \to [n]^t$ that uses*

- *$t$ distinct samples for any $t \in [t_0, n]$, where $t_0 = O(\frac{1}{\theta^2}\log(1/\gamma))$, and*
- *$r = \log(n/t) + \log(1/\gamma)\mathrm{poly}(1/\theta)$ random bits.*

We can now prove Theorem 7.5.

*Proof of Theorem 7.5.* Given the source $X$ and the seed $R$, we construct the extractor $\mathsf{IExt}$ as follows. Set $\delta = 0.9$ and $\tau = 0.1$. Set $\mu = (\delta - 2\tau)/\log(1/\tau)$, $\theta = \tau/\log(1/\tau)$ and $\gamma = \epsilon/4$. Now by Theorem 7.8 there is an explicit $(\mu, \theta, \gamma)$ averaging sampler $\mathsf{Samp} : \{0,1\}^r \to [n]^t$ that uses $t$ distinct samples for any $t \in [t_0, n]$, where $t_0 = O(\frac{1}{\theta^2}\log(1/\gamma)) = O(\log(1/\epsilon))$ and $r = \log(n/t) + \log(1/\gamma)\mathrm{poly}(1/\theta) = \log n + O(\log(1/\epsilon))$. We will set $t = 0.9d + 1$ and $r = 0.1d$. Note that by setting the hidden constant in $d = O(\log(n/\epsilon))$ to be big enough and $\alpha$ to be small enough we can ensure that $0.9d + 1 \in [t_0, n]$, and $r \leq 0.1d$. Thus such a sampler can indeed be constructed.

We now take a slice of $0.1d$ bits from $R$ and let $R = (R_1, R_2)$, where $R_2$ has $0.9d$ bits. We use $R_1$ to sample $t = 0.9d + 1$ distinct bits from $X$, and let the sampled bits be $X'$. By Theorem 7.7 we know that $(R_1, X')$ is $\epsilon/4 + 2^{-\Omega(n)}$-close to $(R_1, W)$ where conditioned on any fixing of $R_1$, $W$ is a $0.6t \geq 0.5d$ source. We will now proceed as if $(R_1, X')$ is $(R_1, W)$, since this only adds error $\epsilon/4 + 2^{-\Omega(n)}$.

Next we fix $R_1$, and note that conditioned on this fixing, $X'$ is a deterministic function of $X$, and thus independent of $R_2$. Further $X'$ has entropy rate 0.6. We now take $R_2$ and let $R_2'$ be $R_2$ padding with a 1 at the end, thus $R_2'$ also has $t = 0.9d + 1$ bits and has min-entropy $0.9d$. Finally we compute the output $\mathsf{IExt}(X, R)$ to be the last $0.3d$ bits of $R_2' \cdot X'$, where the operation is in the field $\mathbb{F}_{2^t}$. By the leftover hash lemma we know that

$$|(\mathsf{IExt}(X, R), R_2) - (U, R_2)| \leq 2 \cdot 2^{-0.1d}.$$

Since $r \leq 0.1d$ we have that $2^{-0.1d} < \gamma = \epsilon/4$. Since conditioned on the fixing of $R_2$ we have that $\mathsf{IExt}(X, R)$ is a deterministic function of $X$, by adding back all the errors we get

$$|(\mathsf{IExt}(X, R), R) - (U, R)| \leq \epsilon/4 + \epsilon/4 + 2^{-\Omega(n)}.$$

By setting $\alpha$ to be small enough we can ensure the total error is at most $\epsilon$, thus we have

$$|(\mathsf{IExt}(X, R), R) - (U, R)| \leq \epsilon.$$

Moreover, for any fixing of the seed $R = r$, the function $\mathsf{IExt}(X, r)$ is a linear function in $X$ because it first selects $t$ bits from $X$ and then performs the operation $R_2' \cdot X'$, which is a linear function since the field is $\mathbb{F}_{2^t}$. Finally, the pre-image size for any fixed seed is the same since first, the pre-image size of $X'$ is always $2^{t-0.3d}$ because $R_2'$ is a fixed non-zero field element, and then given $X'$ to get $X$ it is enough to put any bits for the un-sampled part of $X$. □

We now have the following construction. Let $(X, Y)$ be two independent $(n, (1-\gamma)n)$ source.

- Let $0 < \gamma < \alpha < \beta < 1/70$ be two constants to be chosen later.
- Let $\mathsf{IP}$ be the inner product two-source extractor from Theorem 2.10.
- Let $\mathsf{AdvCB}$ be the correlation breaker with advice from Lemma 5.3 with error $\epsilon = 2^{-\Omega(n/\log n)}$.
- Let $\mathsf{IExt}$ be the invertible linear seeded extractor form Theorem 7.5.

1. Let $n_1 = \alpha n$. Divide $X$ into $X = (X_1, X_2)$ such that $X_1$ has $n_1$ bits and $X_2$ has $n_2 = (1-\alpha)n$ bits. Similarly divide $Y$ into $Y = (Y_1, Y_2)$ such that $Y_1$ has $n_1$ bits and $Y_2$ has $n_2 = (1-\alpha)n$ bits.
2. Compute $Z = \mathsf{IP}(X_1, Y_1)$ which outputs $r = \Omega(n) \leq \alpha n/2$ bits.
3. Let $\mathbb{F}$ be the finite field $\mathbb{F}_{2^{\log n}}$. Let $n_0 = \frac{n_2}{\log n}$. Let $\mathsf{RS} : \mathbb{F}^{n_0} \to \mathbb{F}^n$ be the Reed-Solomon code encoding $n_0$ symbols of $\mathbb{F}$ to $n$ symbols in $\mathbb{F}$ (we slightly abuse the use of $\mathsf{RS}$ to denote both the code and the encoder). Thus $\mathsf{RS}$ is a $[n, n_0, n - n_0 + 1]_n$ error correcting code. Let $X_2'$ be $X_2$ written backwards, and similarly $Y_2'$ be $Y_2$ written backwards. Let $\overline{X}_2 = \mathsf{RS}(X_2')$ and $\overline{Y}_2 = \mathsf{RS}(Y_2')$.
4. Use $Z$ to sample $r/\log n$ distinct symbols from $\overline{X}_2$ (i.e., use each $\log n$ bits to sample a symbol), and write the symbols as a binary string $\tilde{X}_2$. Note that $\tilde{X}_2$ has $r$ bits. Similarly, use $Z$ to sample $r/\log n$ distinct symbols from $\overline{Y}_2$ and obtain a binary string $\tilde{Y}_2$ with $r$ bits.
5. Let $\tilde{\alpha} = X_1 \circ Y_1 \circ \tilde{X}_2 \circ \tilde{Y}_2$. Divide $X_2$ into $X_2 = (X_3, X_4, X_5)$ such that $X_3$ has $n_3 = \beta n$ bits, $X_4$ has $n_4 = 30\beta n$ bits and $X_5$ has $n_5 = (1-\alpha-31\beta)n$ bits. Similarly divide $Y_2 = (Y_3, Y_4, Y_5)$ such that $Y_3$ has $n_3$ bits, $Y_4$ has $n_4$ bits and $Y_5$ has $n_5$ bits.
6. Compute $V = \mathsf{AdvCB}(X_3, Y_3, \tilde{\alpha})$ which outputs $d = n_3/10 = \beta n/10$ bits.
7. Finally compute $W = \mathsf{IExt}(Y_4, V)$ which outputs $\Omega(d) < d/2$ bits.

We now have the following theorem.

**Theorem 7.9.** *There exists a constant $0 < \gamma < 1$ and a non-malleable two-source extractor for $(n, (1-\gamma)n)$ sources with error $2^{-\Omega(n/\log n)}$ and output length $\Omega(n)$.*

*Proof.* We show that the above construction is such a non-malleable two-source extractor. As usual, we will use letters with prime to denote random variables produced from $(X', Y')$. Without loss of generality we assume that $X \neq X'$. The case where $Y \neq Y'$ can be handled in the same way by symmetry.

First we argue that with probability $1 - 2^{-\Omega(n/\log n)}$ over $\tilde{\alpha}, \tilde{\alpha'}$, we have that $\tilde{\alpha} \neq \tilde{\alpha'}$. To see this, note that if $X_1 \neq X_1'$ or $Y_1 \neq Y_1'$ then $\tilde{\alpha} \neq \tilde{\alpha'}$. Otherwise, since $X \neq X'$ we must have $X_2 \neq X_2'$. Thus by the property of the $\mathsf{RS}$ code we know that $\overline{X}_2$ and $\overline{Y}_2$ must differ in at least $n - n_0 > 0.9n$ symbols. Also, since $X_1 = X_1'$ and $Y_1 = Y_1'$ we have $Z = Z'$. Now if $\alpha \geq 3\gamma$ then both $X_1$ and $Y_1$ has min-entropy rate at least $2/3$, thus by Theorem 2.10 we know that

$$(Z, X_1) \approx_{2^{-\Omega(n)}} (U_r, X_1).$$

We can now fix $X_1$, and conditioned on this fixing $Z$ is a deterministic function of $Y$, thus independent of $X_2$. Therefore now we can use $Z$ to sample from $\overline{X}_2$. If $Z$ is uniform then by a Chernoff bound we know that

$$\Pr[\tilde{X}_2 \neq \tilde{X'}_2] \geq 1 - 2^{-r/\log n} = 1 - 2^{-\Omega(n/\log n)}.$$

Thus the total probability that $\tilde{\alpha} \neq \tilde{\alpha'}$ is at least $1 - 2^{-\Omega(n/\log n)} - 2^{-\Omega(n)} = 1 - 2^{-\Omega(n/\log n)}$.

Moreover, by choosing $\alpha < \beta/50$, we can ensure that $r \leq \alpha n < \beta n/50$. Now by Lemma 2.12 we know that conditioned on the fixing of $(\tilde{\alpha}, \tilde{\alpha'})$, with probability $1-2^{-\Omega(n)}$, we have that $H_\infty(X_3) \geq \beta n - \gamma n - \alpha n - 3r \geq 0.9\beta n$ and similarly $H_\infty(Y_3) \geq 0.9\beta n$. Moreover $(X, X')$ and $(Y, Y')$ are still independent.

Now we will use Lemma 5.3. Note that the length of the advice string is $a = 2\alpha n + 2r \leq 3\alpha n$, and $X_3, Y_3$ each has $\beta n$ bits. Thus by choosing the error $\epsilon = 2^{-\Omega(n/\log n)}$ appropriately we can ensure that

$$\beta n \geq C \log a \log(\beta n a/\epsilon),$$

where $C$ is the constant in Lemma 5.3. When this condition holds, by Lemma 5.3 we have that

$$(V, V', Y_3, Y_3') \approx_\epsilon (U_d, V', Y_3, Y_3').$$

We now fix $(Y_3, Y_3')$. Note that conditioned on this fixing, $(V, V')$ is a deterministic function of $(X, X')$, and thus independent of $(Y, Y')$. Moreover the average conditional min-entropy of $Y_4$ is at least $n_4 - \gamma n - \alpha n - 2r - \beta n \geq n_4 - 2\alpha n - \beta n$. Note that $n_4 = 30\beta n$. Thus by choosing $\alpha < \beta/50$ we can ensure that (by Lemma 2.12) with probability $1 - 2^{-\Omega(n)}$, $Y_4$ has min-entropy rate at least 0.95.

Now we can fix $V'$ and then $W'$. Note that conditioned on this fixing, $V$ is still close to uniform, and independent of $Y_4$. Furthermore since the length of $W'$ is at most $d/2 = \beta n/20$, again by Lemma 2.12 we have that with probability $1 - 2^{-\Omega(n)}$, $Y_4$ has min-entropy rate at least 0.9. Thus now by Theorem 7.5 we have that

$$(W, V) \approx_{2^{-\Omega(n)}} (U, V).$$

Note that conditioned on the fixing of $V$, $W$ is a deterministic function of $Y$. Since we have already fixed $(V', W')$, by adding back all the errors we get that

$$(W, W', X, X') \approx_{2^{-\Omega(n/\log n)}} (U, W', X, X').$$

□

## 7.2 Efficiently sampling algorithm and the non-malleable code

We now show that given an output of the non-malleable two-source extractor, we can efficiently uniformly sample from the pre-image of this output. First we have the following main lemma.

**Lemma 7.10.** *Given any arbitrary fixing of $(X_1, \tilde{X}_2, X_3, Y_1, \tilde{Y}_2, Y_3, W)$, there is an efficient procedure to uniformly sample from the pre-image $(X, Y)$. Moreover, for any fixing of $(X_1, \tilde{X}_2, X_3, Y_1, \tilde{Y}_2, Y_3, W)$, the pre-image has the same size.*

*Proof.* Assume that we are given $(X_1, \tilde{X}_2, X_3, Y_1, \tilde{Y}_2, Y_3, W) = (x_1, \tilde{x}_2, x_3, y_1, \tilde{y}_2, y_3, w)$ for arbitrary $(x_1, \tilde{x}_2, x_3, y_1, \tilde{y}_2, y_3, w)$. We need to sample from the corresponding $(X_4, X_5, Y_4, Y_5)$. First we can compute $z = \mathsf{IP}(x_1, y_1)$ which tells us what symbols of the $\mathsf{RS}$ codes are sampled. Next, we can compute $v = \mathsf{AdvCB}(x_3, y_3, \tilde{\alpha})$ where $\tilde{\alpha} = x_1 \circ y_1 \circ \tilde{x}_2 \circ \tilde{y}_2$. Now note that $W = \mathsf{IExt}(Y_4, V)$, therefore by Theorem 7.5 we can efficiently and uniformly sample the pre-image of $w$, which is $Y_4$, by inverting a system of linear equations. Also, Theorem 7.5 guarantees that for any $(v, w)$ the pre-image has the same size.

Now once we have sampled $Y_4 = y_4$, we will continue to sample $(X_4, X_5, Y_5)$. Since these are different bits in $(X, Y)$ than the bits we have already obtained, they can almost be sampled arbitrarily, except they need to satisfy the linear constraints imposed by the $\mathsf{RS}$ codes: $\tilde{Y}_2 = y_2$ and $\tilde{X}_2 = x_2$. We first look at the $Y$ part. Note that $\tilde{Y}_2 = y_2$ gives us $r/\log n \leq \alpha n/(2\log n) < n/(4\log n)$ equations in the field $\mathbb{F}_{2^{\log n}}$. Also note that now $(Y_1, Y_3, Y_4)$ are fixed and $Y_5$ are the variables. Since the length of $Y_5$ is $n_5 = n - \alpha n - \beta n - 30\beta n > n/2$ (as $\beta < 1/70$), this gives us at least $n/(2\log n)$ variables in the field $\mathbb{F}_{2^{\log n}}$. Finally, note that when we encode $Y_2$ using the $\mathsf{RS}$ code, we encode it as $\mathsf{RS}(Y_2')$ where $Y_2'$ is $Y_2$ written backwards. Thus the coefficient matrix of the equations with variables in $Y_5$ is

$$G = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_s \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^t & \alpha_2^t & \cdots & \alpha_s^t \end{pmatrix}$$

where $s = r/\log n$, $t = n_5/\log n$, and $\alpha_1, \ldots, \alpha_s$ are distinct field elements of $\mathbb{F}_{2^{\log n}}$.

Note that $t = n_5/\log n > n/(2\log n) > r/\log n = s$, thus all the columns in the matrix are linearly independent, and the kernel of the matrix has dimension exactly $t - s$ for any $\alpha_1, \ldots, \alpha_s$. Therefore, we can efficiently sample $Y_5$ by inverting the system of linear equations, and moreover for any fixing of $(Y_1, Y_3, Y_4, Z) = (y_1, y_3, y_4, z)$ the pre-image always has the same size.

The argument for sampling the $X$ part is exactly the same, except now $X$ has more variables ($(X_4, X_5)$) than $Y$. □

We now have the following main theorem.

**Theorem 7.11.** *Given any output $W = w$ of the non-malleable two-source extractor, there is an efficient procedure to uniformly sample from the pre-image $(X, Y)$.*

*Proof.* The sampling procedure is as follows. We first uniformly randomly generate $(X_1, \tilde{X}_2, X_3, Y_1, \tilde{Y}_2, Y_3) = (x_1, \tilde{x}_2, x_3, y_1, \tilde{y}_2, y_3)$, then we use Lemma 7.10 to generate $(X, Y)$. By Lemma 7.10, for any fixing of $(X_1, \tilde{X}_2, X_3, Y_1, \tilde{Y}_2, Y_3, W)$, the pre-image has the same size. Thus indeed this procedure uniformly samples from the pre-image $(X, Y)$. □

Combining Theorem 7.4, Theorem 7.9, and Theorem 7.11, we immediately obtain the following theorem.

**Theorem 7.12.** *For any $n \in \mathbb{N}$ there exists an explicit non-malleable code with efficient encoder/decoder in the 2-split-state model with block length $2n$, rate $\Omega(1/\log n)$ and error $= 2^{-\Omega(n/\log n)}$.*

# 8 $t$-Non-Malleable Extractors and Applications to Independent Source Extractors

In this section, we extend our results to the case of $t$ tampering functions, and use them to obtain improved results of independent source extractors.

We first prove that any $s$-source non-malleable extractor with sufficiently small error must be a strong $s$-source non-malleable extractor. Formally, we have

**Theorem 8.1.** *Suppose* $\mathsf{nmExt} : (\{0,1\}^n)^s \to \{0,1\}^m$ *is an* $s$*-source non-malleable extractor with error* $\epsilon$ *for min-entropy* $k$*. Then for any* $k' \geq k$, $\mathsf{nmExt}$ *is a strong* $s$*-source non-malleable extractor for min-entropy* $k'$ *with error* $2^{2m}(\epsilon + 2^{k+1-k'})$*.*

*Proof.* Let $X_1, \cdots, X_s$ be independent $(n, k')$ sources and $X'_1 = f_1(X_1), \cdots, X'_s = f_s(X_s)$ where for each $i$, $f_i : \{0,1\}^n \to \{0,1\}^n$ is a deterministic function such that at least one of them has no fixed point. Consider any $i$. Let $X_{<i} = (X_1, \cdots, X_{i-1})$, $X_{>i} = (X_{i+1}, \cdots, X_s)$ and similarly $X'_{<i} = (X'_1, \cdots, X'_{i-1})$, $X'_{>i} = (X'_{i+1}, \cdots, X'_s)$. Now for any $(z, z') \in (\{0,1\}^m)^2$, define the set of bad $y$'s for $(z, z')$ to be

$$B_{z,z'} = \{y : \left|\Pr[\mathsf{nmExt}(X_{<i}, y, X_{>i}) = z, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z'] - 2^{-m}\Pr[\mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z']\right| > \epsilon.\}$$

We have the following claim.

**Claim 8.2.** *For any* $(z, z')$*, we have* $|B_{z,z'}| < 2^{k+1}$*.*

Suppose not, then define

$$B^+_{z,z'} = \{y : \Pr[\mathsf{nmExt}(X_{<i}, y, X_{>i}) = z, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z'] - 2^{-m}\Pr[\mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z'] > \epsilon.\}$$

and

$$B^-_{z,z'} = \{y : \Pr[\mathsf{nmExt}(X_{<i}, y, X_{>i}) = z, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z'] - 2^{-m}\Pr[\mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z'] < -\epsilon.\}$$

We have that either $|B^+_{z,z'}| \geq 2^k$ or $|B^-_{z,z'}| \geq 2^k$. Without loss of generality assume that $|B^+_{z,z'}| \geq 2^k$. Then, let $Y$ be the uniform distribution over $B^+_{z,z'}$. We have that $Y$ is an $(n, k)$ source, but

$$\begin{aligned}
&\Pr[(\mathsf{nmExt}(X_{<i}, Y, X_{>i}, \mathsf{nmExt}(X'_{<i}, f_i(Y), X'_{>i})) = (z, z')] - \Pr[(U_m, \mathsf{nmExt}(X'_{<i}, f_i(Y), X'_{>i})) = (z, z')] \\
=& \sum_{y \in B^+_{z,z'}} \Pr[Y = y]\Pr[(\mathsf{nmExt}(X_{<i}, y, X_{>i}, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i})) = (z, z')] \\
&- 2^{-m} \sum_{y \in B^+_{z,z'}} \Pr[Y = y]\Pr[\mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z'] \\
=& \sum_{y \in B^+_{z,z'}} \Pr[Y = y](\Pr[(\mathsf{nmExt}(X_{<i}, y, X_{>i}, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i})) = (z, z')] \\
&- 2^{-m}\Pr[\mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}) = z']) \\
>& \epsilon,
\end{aligned}$$

which contradicts the fact that $\mathsf{nmExt}$ is a non-malleable extractor.

Now let $B = \cup_{z,z'} B_{z,z'}$, we have that $|B| \leq 2^{2m} 2^{k+1}$. Thus, we now have that

$$\begin{aligned}
&\left|(\mathsf{nmExt}(X_1, \cdots, X_s), \mathsf{nmExt}(X'_1, \cdots, X'_s), X_i, X'_i) - (U_m, \mathsf{nmExt}(X'_1, \cdots, X'_s), X_i, X'_i)\right| \\
=& \sum_{y \in \{0,1\}^n} \Pr[X_i = y] \left|(\mathsf{nmExt}(X_{<i}, y, X_{>i}, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i})) - (U_m, \mathsf{nmExt}(X'_{<i}, f_i(y), X'_{>i}))\right| \\
\leq& \Pr[X_i \in B] \cdot 1 + \Pr[X_i \notin B] 2^{2m} \epsilon \\
\leq& 2^{2m}(\epsilon + 2^{k+1-k'})
\end{aligned}$$

□

We now have the following lemma.

**Lemma 8.3.** *Suppose that there exists a constant $\gamma > 0$ and an explicit construction of a strong non-malleable $s$-source extractor $\mathsf{nmExt} : (\{0,1\}^n)^s \to \{0,1\}^m$ for $(n, (1-2\gamma)n)$ sources which outputs $\Omega(n)$ bits with error $2^{-\Omega(n)}$. Then given any $t \in \mathbb{N}$ there is an explicit function $\mathsf{AdvCB} : (\{0,1\}^n)^s \times \{0,1\}^a \to \{0,1\}^m$ with $m = \Omega(a)$ and the following property.*

*Let $X_1, \cdots, X_s$ be $s$ independent uniform strings on $n$ bits, and $\alpha, \alpha_1, \cdots, \alpha_t$ be $t+1$ strings on $a$ bits such that $\forall j \in [t], \alpha \neq \alpha_j$. Let $X_i^j, i \in [s], j \in [t]$ be random variables on $n$ bits such that $(\overline{X}_1 = (X_1, \{X_1^j, j \in [t]\}), \cdots, \overline{X}_s = (X_s, \{X_s^j, j \in [t]\}))$ are independent (i.e., each $X_i^j$ only depends on $X_i$). Let $Z = \mathsf{AdvCB}(X_1, \cdots, X_s, \alpha)$ and $Z^j = \mathsf{AdvCB}(X_1^j, \cdots, X_s^j, \alpha_j)$ for any $j \in [t]$. Then as long as $n \geq 2(t+1)^2 a/\gamma$, we have that $\forall i \in [s]$,*

$$\left|(Z, \{Z^1, \cdots, Z^t\}, X_i) - (U_m, \{Z^1, \cdots, Z^t\}, X_i)\right| \leq ts2^{-\Omega(a)}.$$

We construct the function $\mathsf{AdvCB}$ as follows. Let $\mathsf{Ext}$ be an optimal seeded extractor from Theorem 2.9 that uses $O(\log(n/\epsilon))$ bits to extract from an $(n,k)$ source and output $0.9k$ bits.

1. $\forall i \in [s]$, let $V_i$ be a slice of $X_i$ with length $a/\gamma$.
2. Repeat the following step for $t$ times: $\forall i \in [s]$, let $\tilde{V}_i = V_i \circ \alpha$. Compute $R = \mathsf{nmExt}(\tilde{V}_1, \cdots, \tilde{V}_s)$. Then $\forall i \in [s]$, compute $V'_i = \mathsf{Ext}(X_i, R)$ and outputs $a/\gamma$ bits. Finally $\forall i \in [s]$, let $V_i = V'_i$.
3. Output $R$ from the last step, i.e., the computation of $V'_i = \mathsf{Ext}(X_i, R)$ and $V_i = V'_i$ in the above iteration can be omitted for the $t$'th execution.

We now prove the lemma.

*Proof.* We prove the function $\mathsf{AdvCB}$ described above is the desired function. We will use letters with superscript $j$ to denote random variables produced from $(X_i^j, i \in [s])$ and $\alpha_j$. By fixing additional randomness, without loss of generality we can assume that $\forall i \in [s]$, we have that $\forall j \in [t]$, $X_i^j$ is a deterministic function of $X_i$. We will use induction to prove the following claim.

**Claim 8.4.** *At the beginning of the $\ell$'th iteration, conditioned on the fixing of previous random variables (produced in previous rounds), we have that*

- *$X_1, \cdots, X_s$ are still independent.*

- $\forall j \in [t]$, *$V_i$ is a deterministic function of $X_i$ and $V_i^j$ is a deterministic function of $X_i^j$.*
- $\forall i \in [s]$, *the average conditional min-entropy of $X_i$ is at least $n - (\ell-1)(t+1)a/\gamma$.*

*At the end of the $\ell$'th iteration, we have that $\forall i$ and any $S \subseteq [t]$ with $|S| = \ell$,*

$$\left|(R, \{R^j, j \in S\}, V_i, \{V_i^j, j \in S\}) - (U_m, \{R^j, j \in S\}, V_i, \{V_i^j, j \in S\})\right| \leq \ell s 2^{-\Omega(a)}.$$

To prove the claim, first note that since $\mathsf{nmExt}$ is a strong non-malleable $s$-source extractor $(n, (1-2\gamma)n)$ sources with error $2^{-\Omega(n)}$, it is also a strong non-malleable $s$-source extractor for average conditional min-entropy $(1-\gamma)n$ with error $2^{-\Omega(n)}$, by Lemma 2.6.

For the base case where $\ell = 1$, clearly at the beginning of the first iteration, $X_1, \cdots, X_s$ are independent. Further, $\forall i \in s$, $V_i$ is a deterministic function of $X_i$ and $V_i^j$ is a deterministic function of $X_i^j$, $\forall j \in [t]$. Also the min-entropy of each $X_i$ is at least $n$. Now note that each $\tilde{V}_i$ has min-entropy rate at least $(a/\gamma)/(a + a/\gamma) > 1 - \gamma$, and $\tilde{V}_i \neq \tilde{V}_i^j$ for any $j \in [t]$. Thus the claim follows by the assumption that $\mathsf{nmExt}$ is a strong non-malleable $s$-source extractor.

We next assume the claim holds for $\ell$ and show that it holds for $\ell+1$. The first three properties can be directly verified. We now prove the last property. Consider any set $S \subseteq [t]$ with $|S| = \ell + 1$. Pick any $j_0 \in S$ and let $S' = S \setminus \{j_0\}$. By the claim we know that at the end of iteration $\ell$, we have that $\forall i \in [s]$,

$$\left|(R, \{R^j, j \in S'\}, V_i, \{V_i^j, j \in S'\}) - (U_m, \{R^j, j \in S'\}, V_i, \{V_i^j, j \in S'\})\right| \leq \ell s 2^{-\Omega(a)}.$$

Consider any $i \in [s]$. We now fix $(V_i, \{V_i^j, j \in S'\})$. Note that all these random variables are deterministic functions of $(X_i, \{X_i^j, j \in S'\})$, which are in turn deterministic functions of $X_i$. Thus conditioned on this fixing, $X_1, \cdots, X_s$ are still independent. Also note that conditioned on this fixing, $(R, \{R^j, j \in S'\})$ is a deterministic function of $(\{V_h, h \neq i\}, \{V_h^j, h \neq i, j \in S'\})$, and therefore independent of $X_i$ and its derived random variables. Thus, we can further fix all the remaining $\{V_i^j, j \in [t]\}$ without affecting $(R, \{R^j, j \in S'\})$. Note that now the average conditional min-entropy of $X_i$ is at least $n - \ell(t+1)a/\gamma - (t+1)a/\gamma = n - (\ell+1)(t+1)a/\gamma$.

Now we have that $R$ is still close to uniform given $\{R^j, j \in S'\}$. We now fix all $\{R^j, j \in S'\}$ and then all $\{V_i'^j = \mathsf{Ext}(X_i, R^j), j \in S'\}$. Note that fixing $\{R^j, j \in S'\}$ does not affect $X_i$, and conditioned on the fixing of all $\{R^j, j \in S'\}$, we have that $\{V_i'^j = \mathsf{Ext}(X_i, R^j), j \in S'\}$ is a deterministic function of $(X_i, \{X_i^j, j \in S'\})$, which are in turn deterministic functions of $X_i$. Now the average conditional min-entropy of $X_i$ is at least $n-(\ell+1)(t+1)a/\gamma - \ell a/\gamma \geq n-(t+1)^2 a/\gamma > n/2$. Thus by Theorem 2.9 (and noticing that $R$ is still close to uniform and independent of $X_i$) we have

$$|(V_i', R) - (U_m, R)| \leq 2^{-\Omega(a)}.$$

Note that given $R$, $V_i'$ is again a deterministic function of $X_i$. Thus (ignoring the $\ell s 2^{-\Omega(a)}$ for now) we have the following inequality.

$$\left|(V_i', R, \{R^j, j \in S'\}, \{V_i'^j, j \in S'\}, \{V_i^j, j \in [t]\}) - (U_m, R, \{R^j, j \in S'\}, \{V_i'^j, j \in S'\}, \{V_i^j, j \in [t]\})\right| \leq 2^{-\Omega(a)}.$$

Furthermore, conditioned on the fixing of $(R, \{R^j, j \in S'\}, \{V_i'^j, j \in S'\}, \{V_i^j, j \in [t]\})$, $V_i'$ is a deterministic function of $X_i$ and therefore independent of $\{X_h, h \neq i\}$. Thus, we can also fix all the other $\{V_h^j, j \in [t], h \neq i\}$ without affecting the inequality. Thus we obtain the following.

$$\left|(V_i', \{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\}) - (U_m, \{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\})\right| \leq 2^{-\Omega(a)}.$$

Using the same argument, we can also show that conditioned on the fixing of $(\{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\})$, $V_i'^{j_0}$ is a deterministic function of $X_i^{j_0}$, which in turn is a deterministic function of $X_i$. However, we don't know if $V_i'^{j_0}$ is close to uniform, and it may be correlated with $V_i'$.

We can repeat the above argument for any $i \in [s]$, thus we obtain the following conclusion.

- $\forall i \in [s]$, we have

$$\left|(V_i', \{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\}) - (U_m, \{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\})\right| \leq 2^{-\Omega(a)}.$$

- Further, $\forall i \in [s]$, conditioned on the fixing of $(\{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\})$, we have that $(V_i', V_i'^{j_0})$ is a deterministic function of $X_i$.

Now fix $(\{V_i'^j, j \in S'\}, \{V_h^j, j \in [t], h \in [s]\})$. Note that conditioned on this fixing, $X_1, \cdots, X_s$ are still independent. Thus $(V_i', V_i'^{j_0})$ are also independent. By the fact that $\mathsf{nmExt}$ is a strong non-malleable $s$-source extractor, we have that $\forall i \in [s]$,

$$\left|(R, R^{j_0}, V_i', V_i'^{j_0}) - (U_m, R^{j_0}, V_i', V_i'^{j_0})\right| \leq 2^{-\Omega(a)}.$$

Since we have fixed all the $(\{V_i'^j, j \in S'\})$ before, and each new $(R^j, j \in S')$ is now a deterministic function of $(\{V_i'^j, j \in S'\})$, by adding back all the errors we also have that

$$\left|(R, \{R^j, j \in S\}, V_i', \{V_i'^j, j \in S\}) - (U_m, \{R^j, j \in S\}, V_i', \{V_i'^j, j \in S\})\right| \leq (\ell + 1)s2^{-\Omega(a)}.$$

Note that at the end of iteration we replace $V_i$ with $V_i'$, so the claim holds and the theorem is proved. □

We now have the following theorem.

**Theorem 8.5.** *Suppose there is a constant $\gamma > 0$ and an explicit non-malleable $(s+1)$-source extractor for $(n, (1-\gamma)n)$ sources with error $2^{-\Omega(n)}$ and output length $\Omega(n)$. Then there is a constant $C > 0$ such that for any $0 < \epsilon < 1$ with $k \geq Ct^2 \log(n/\epsilon)$, there is an explicit strong seeded $t$-non-malleable extractor for $s$ independent $(n, k)$ sources with seed length $d = Ct^2 \log(n/\epsilon)$, error $O(ts\epsilon)$ and output length $\Omega(\log(1/\epsilon))$.*

The construction of the seeded non-malleable extractor for $s$ independent $(n, k)$ sources is as follows. Let the sources be $X_1, \cdots, X_s$ and the seed be $Y$.

- Let $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^{d'} \to \{0,1\}^m$ be an optimal seeded extractor from Theorem 2.9, which uses $d_1 = O(\log(n/\epsilon))$ random bits to extract from $(n,k)$ sources and output $m = 0.9k$ bits.
- Let $\mathsf{Ext}_1, \mathsf{Ext}_2$ be optimal seeded extractors from Theorem 2.9.
- Let $\mathsf{IP}$ be the inner product two-source extractor from Theorem 2.10.
- Let $\mathsf{AdvCB}$ be the correlation breaker with advice from Lemma 8.3.
- Let $\mathsf{AdvGen}$ be the advice generator from Theorem 6.1.

1. Take a small slice $Y_1$ of $Y$ with length $d_1 = O(\log(n/\epsilon))$, for every $i \in [s]$, compute $Z_i = \mathsf{Ext}'(X_i, Y_1)$ which outputs $0.9k$ bits.
2. Use $Z_1$ and $X$ to compute $\mathsf{AdvGen}(X, Y)$. Specifically, as in Theorem 6.1, take a small slice $\overline{Z}_1$ of $Z_1$ with length $d_2 = O(\log(n/\epsilon))$ and compute $Y_2 = g(Y, \overline{Z}_1)$ which outputs $d_3 = O(\log(1/\epsilon))$ bits. Let $\mathsf{AdvGen}(X_1, Y) = (Y_1, Y_2) = \alpha$. Note that the length of the advice is $a = d_1 + d_3 = O(\log(n/\epsilon))$. We choose the hidden constant to be big enough such that the term $2^{-\Omega(a)}$ in Lemma 8.3 is at most $\epsilon$.
3. Let $d_4 = max\{d_2, d_1 + d_3\}$. Take a slice of $Y_3$ of $Y$ with length $d_5 = 3(t+1)d_4$, and a slice $Z_3$ of $Z$ with length $d_5 = 3td_4$. Compute $R = \mathsf{IP}(Y_3, Z_3)$.
4. Compute $\tilde{Y} = \mathsf{Ext}_1(Y, R)$ which outputs $m_1 = 0.5d$ bits, and $\tilde{Z}_1 = \mathsf{Ext}_2(Z_1, R)$ which outputs $m_1$ bits. For $i = 2, \cdots, s$, truncate each $Z_i$ to $\tilde{Z}_i$ with $m_1$ bits.
5. Output $V = \mathsf{AdvCB}(\tilde{Y}, \tilde{Z}_1, \cdots, \tilde{Z}_s, \alpha)$.

*Proof.* Suppose we have $t$ tampered seeds $Y^j = f_j(Y), j \in [t]$, where each $f_j$ has no fixed points. We will use letters with superscript $j$ to denote random variables obtained from $(Y^j, X_1, \cdots, X_s)$. First, by Theorem 2.9, we have that for any $i \in [s]$,

$$|(Z_i, Y_1) - (U, Y_1)| \leq \epsilon.$$

Since conditioned on the fixing of $Y_1$, each $Z_i$ is a deterministic function of $X_i$ and thus independent, we have

$$|(Z_1, \cdots, Z_s, Y_1) - (U, \cdots, U, Y_1)| \leq s\epsilon.$$

We will now proceed as if $(Z_1, \cdots, Z_s)$ are uniform and independent, given $Y_1$. Take any $j \in [t]$, by Theorem 6.1, we know that with probability $1 - \epsilon$ over the fixing of $(Y_1, \overline{Z}_1, Y_2, Y_1^j, \overline{Z}_1^j, Y_2^j)$, we have $\alpha \neq \alpha^j$. Thus, with probability $1 - t\epsilon$ over the fixing of $H = (Y_1, \overline{Z}_1, Y_2, \{Y_1^j, \overline{Z}_1^j, Y_2^j, j \in [t]\})$, we have that $\forall j, \alpha \neq \alpha^j$. Furthermore, notice that conditioned on the fixing of $H$, we have that $(Y, \{Y^j, j \in [t]\}$ and $(Z_1, \{Z_1^j, j \in [t]\})$ are still independent, the average conditional min-entropy of $Y_3$ is at least $d_5 - (t+1)(d_1 + d_3) \geq 2(t+1)d_4$, and the average conditional min-entropy of $Z_3$ is at least $d_5 - (t+1)d_2 \geq 2(t+1)d_4$. Also note that the fixing of $H$ does not affect $Z_2, \{Z_2^j, j \in [t]\}, \cdots, Z_s, \{Z_s^j, j \in [t]\}$.

Now by Theorem 2.10, we have that

$$|(R, Y_3) - (U, Y_3)| \leq \epsilon.$$

Note that conditioned on the fixing of $(Y_3, \{Y_3^j, j \in [t]\})$, $(R, \{R^j, j \in [t]\})$ is a deterministic function of $(Z_3, \{Z_3^j, j \in [t]\})$, and thus independent of $(Y, \{Y^j, j \in [t]\})$. Moreover $R$ is close to uniform and the average conditional min-entropy of $Y$ is at least $d - (t+1)(d_1 + d_3 + d_5) = d - O(t^2 \log(n/\epsilon))$. Thus by taking $C$ to be large enough we have that $d - O(t^2 \log(n/\epsilon)) > 2d/3$. Thus by Theorem 2.9 we have that

$$|(\tilde{Y}, R) - (U, R)| \leq \epsilon.$$

Note that conditioned on the further fixing of $(R, \{R^j, j \in [t]\})$, $(\tilde{Y}, \{\tilde{Y}^j, j \in [t]\})$ is a deterministic function of $(Y, \{Y^j, j \in [t]\})$. Thus we can further fix $(Z_3, \{Z_3^j, j \in [t]\})$ without affecting the above inequality. Similarly, we also have

$$|(R, Z_3) - (U, Z_3)| \leq \epsilon.$$

Note that conditioned on the fixing of $(Z_3, \{Z_3^j, j \in [t]\})$, $(R, \{R^j, j \in [t]\})$ is a deterministic function of $(Y_3, \{Y_3^j, j \in [t]\}$, and thus independent of $(Z_1, \{Z_1^j, j \in [t]\})$. Moreover $R$ is close to uniform and the average conditional min-entropy of $Z_1$ is at least $0.9k - (t+1)(d_2 + d_5) = 0.9k - O(t^2 \log(n/\epsilon))$. Thus by taking $C$ to be large enough we have that $0.9k - O(t^2 \log(n/\epsilon)) > 2d/3$. Thus by Theorem 2.9 we have that

$$|(\tilde{Z}_1, R) - (U, R)| \leq \epsilon.$$

Note that conditioned on the further fixing of $(R, \{R^j, j \in [t]\})$, $(\tilde{Z}_1, \{\tilde{Z}_1^j, j \in [t]\})$ is a deterministic function of $(Z_1, \{Z_1^j, j \in [t]\})$. Thus we can further fix $(Y_3, \{Y_3^j, j \in [t]\})$ without affecting the above inequality. Note that none of these affects $Z_2, \{Z_2^j, j \in [t]\}, \cdots, Z_s, \{Z_s^j, j \in [t]\}$. Therefore, combining the above we have that with probability $1 - O(t\epsilon)$ over the fixing of $\bar{H} = (H, Z_3, \{Z_3^j, j \in [t]\}, R, \{R^j, j \in [t]\}, Y_3, \{Y_3^j, j \in [t]\})$,

- $\forall j, \alpha \neq \alpha^j$.
- $(\tilde{Y}, \{\tilde{Y}^j, j \in [t]\}), (\tilde{Z}_1, \{\tilde{Z}_1^j, j \in [t]\}), \cdots, (\tilde{Z}_s, \{\tilde{Z}_s^j, j \in [t]\})$ are independent.
- $(\tilde{Y}, \tilde{Z}_1, \cdots, \tilde{Z}_s) \approx_{O(s\epsilon)} (U_{m_1}, \cdots, U_{m_1})$.

Next, using Theorem 8.1, we see that the non-malleable $(s+1)$-source extractor is also a strong non-malleable $(s+1)$-source extractor for $(n, (1-\gamma/2)n)$ sources with error $2^{2m'}(2^{-\Omega(n)} + 2^{1-\gamma n/2})$, where $m'$ is the output length of the extractor. By truncating the output if necessary, we can ensure that $m' = \Omega(n)$ and $2^{2m'}(2^{-\Omega(n)} + 2^{1-\gamma n/2}) = 2^{-\Omega(n)}$. Thus the non-malleable $(s+1)$-source extractor is also a strong non-malleable $(s+1)$-source extractor for $(n, (1-\gamma/2)n)$ sources with error $2^{-\Omega(n)}$ and output length $\Omega(n)$.

We now apply Lemma 8.3. First ignoring the error, and note that the length of each $(\tilde{Y}, \tilde{Z}_1, \cdots, \tilde{Z}_s)$ is $0.5d$ where $d = Ct^2 \log(n/\epsilon)$, and the length of advice is $a = O(\log(n/\epsilon))$. By choosing the constant $C$ large enough we can ensure that $0.5d \geq 2(t+1)^2 a/(\gamma/4)$. Therefore by Lemma 8.3, we have that the output has length $\Omega(\log(1/\epsilon))$, and $\forall i$,

$$\left|(V, \{V^1, \cdots, V^t\}, X_i) - (U_m, \{V^1, \cdots, V^t\}, X_i)\right| \leq t(s+1)\epsilon.$$

Adding back all the errors we see that the construction is a seeded $t$-non-malleable extractor for $s$ independent $(n, k)$ sources with error $O(ts\epsilon)$ and output length $\Omega(\log(1/\epsilon))$. □

The above construction and theorem can also be easily generalized to the case where we do not have non-malleable $s+1$-source extractors with asymptotically optimal error. For example, suppose to get error $\epsilon$ the non-malleable $s+1$-source extractor needs $(f(\epsilon),(1-\gamma)f(\epsilon))$ sources for some function $f$ (note that $f(\epsilon)$ is at least $O(\log(1/\epsilon))$, then all we need to change is that in Lemma 8.3, the size of each $V_i$ should become $O(\log n + f(\epsilon))$. Since the length of the advice is always going to be $O(\log(n/\epsilon))$, this ensures that each time when we apply the non-malleable $(s+1)$-source extractor, the sources have entropy rate at least $1-\gamma$ and error $\epsilon$. Now the same analysis in Theorem 8.5 goes through, as long as $k, d \geq Ct^2(\log n + f(\epsilon))$. Thus, we have the following theorem.

**Theorem 8.6.** *Suppose there is a function $f$, a constant $\gamma > 0$ and an explicit non-malleable $(s+1)$-source extractor for $(f(\epsilon),(1-\gamma)f(\epsilon))$ sources with error $\epsilon$ and output length $\Omega(f(\epsilon))$. Then there is a constant $C > 0$ such that for any $0 < \epsilon < 1$ with $k \geq Ct^2(\log n + f(\epsilon))$, there is an explicit strong seeded $t$-non-malleable extractor for $s$ independent $(n,k)$ sources with seed length $d = Ct^2(\log n + f(\epsilon))$, error $O(ts\epsilon)$ and output length $\Omega(f(\epsilon))$.*

The constructions and theorems can also be extended to the case of $t$-non-malleable extractors for $s$ independent sources, we omit the details for now.

We now combine Theorem 8.5 and Theorem 8.6 with known constructions of non-malleable $s$-source extractors to obtain seeded $t$-non-malleable extractors. By combining Theorem 8.6 and Theorem 7.9, we have the following theorem (note that here $f(\epsilon) = O(\log(1/\epsilon)\log\log(1/\epsilon))$.

**Theorem 8.7.** *There exists a constant $C > 1$ such that for any $t \in \mathbb{N}$, $0 < \epsilon < 1$ and $k \geq Ct^2(\log n + \log(1/\epsilon)\log\log(1/\epsilon))$, there is an explicit strong seeded $t$-non-malleable extractor for $(n,k)$ sources with seed length $d = Ct^2(\log n + \log(1/\epsilon)\log\log(1/\epsilon))$, output length $\Omega(\log(1/\epsilon)\log\log(1/\epsilon))$ and error $O(t\epsilon)$.*

Next, we use the following theorem proved by Chattopadhyay and Zuckerman [CZ14].

**Theorem 8.8** ([CZ14])**.** *There is a constant $0 < \gamma < 1$ and an explicit non-malleable 10-source extractor for $(n,(1-\gamma)n)$ sources with error $2^{-\Omega(n)}$ and output length $\Omega(n)$.*

Combining this theorem with Theorem 8.5, we have the following theorem.

**Theorem 8.9.** *There exists a constant $C > 1$ such that for any $t \in \mathbb{N}$, $0 < \epsilon < 1$ and $k \geq Ct^2(\log(n/\epsilon))$, there is an explicit strong seeded $t$-non-malleable extractor for 9 independent $(n,k)$ sources with seed length $d = Ct^2(\log(n/\epsilon))$, output length $\Omega(\log(1/\epsilon))$ and error $O(t\epsilon)$.*

By using improved somewhere random condensers as samplers and following the framework in [CZ16], Ben-Aroya et. al [BADTS16] proved the following theorem.

**Theorem 8.10.** *[BADTS16] Suppose there is a function $f$ and an explicit strong seeded $t$-non-malleable extractor for $s$ independent $(n,k')$ sources with seed length and entropy requirement $d = k' = f(t,\epsilon)$, then there for every constant $\epsilon > 0$ exist constants $t = t(\epsilon), c = c(\epsilon)$ and an explicit extractor $\mathsf{Ext} : (\{0,1\}^n)^s \to \{0,1\}$ for $s$ independent $(n,k)$ sources with $k \geq f(t,1/n^c)$ and error $\epsilon$.*

**Remark 8.11.** The original construction in [BADTS16] is just for two sources, but it extends directly to any $s$ sources just by treating $s-1$ sources as one source.

We can now use above theorems to get improved constructions of independent source extractors. For example, combining the above theorem with Theorem 8.7, we immediately obtain the following theorem.

**Theorem 8.12.** *For every constant $\epsilon > 0$ exists a constant $c > 1$ and an explicit two-source extractor $\mathsf{Ext} : (\{0,1\}^n)^2 \to \{0,1\}$ for min-entropy $k \geq c \log n \log\log n$, with error $\epsilon$.*

Using Theorem 8.9 instead, we obtain the following theorem.

**Theorem 8.13.** *For every constant $\epsilon > 0$ exists a constant $c > 1$ and an explicit ten-source extractor $\mathsf{Ext} : (\{0,1\}^n)^{10} \to \{0,1\}$ for min-entropy $k \geq c \log n$, with error $\epsilon$.*

# 9 Conclusions and Open Problems

Previous work in the literature have established connections between seeded non-malleable extractors and two-source extractors, and connections between non-malleable two-source (or multi-source) extractors and non-malleable codes in the split-state model. In this paper we further established connections between seeded non-malleable extractors and non-malleable two-source extractors. Thus, all these four objects are closely related to each other. Using improved independence preserving mergers, we give improved constructions of seeded non-malleable extractors, two-source extractors, non-malleable two-source extractors and non-malleable codes in the two-split-state model. These constructions are quite close to optimal (in terms of the entropy requirement). Thus, the obvious open problem is to achieve optimal constructions for all of them, i.e., seeded non-malleable extractor with seed length and entropy $O(\log(n/\epsilon))$, non-malleable two-source extractor for entropy $(1-\gamma)n$ with error $2^{-\Omega(n)}$ and output length $\Omega(n)$. In turn, these will give explicit two-source extractors for $O(\log n)$ min-entropy (with one bit output and any constant error), and constant-rate non-malleable codes in the two-split-state model.

On the other hand, all recent constructions of two-source extractors follow the framework of [CZ16], and thus the error is either $1/\text{poly}(n)$ or any constant. So far, negligible error can only be achieved by using three sources [Li15b], or two-sources when the min-entropy is at least $0.49n$ [Bou05]. Constructing two-source extractors with smaller error, for smaller min-entropy is an interesting open problem, and seems to require new ideas.

# References


[ADKO15] D. Aggarwal, Y. Dodis, T. Kazana, and M. Obremski. Non-malleable reductions and applications. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, 2015.

[ADL14] Divesh Aggarwal, Yevgeniy Dodis, and Shachar Lovett. Non-malleable codes from additive combinatorics. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2014.

[Agg14] Divesh Aggarwal. Affine-evasive sets modulo a prime. Technical Report 2014/328, Cryptology ePrint Archive, 2014.

[BADTS16] Avraham Ben-Aroya, Dean Doron, and Amnon Ta-Shma. Explicit two-source extractors for near-logarithmic min-entropy. Technical Report TR16-088, ECCC, 2016.

[BBR88] Charles H. Bennett, Gilles Brassard, and Jean-Marc Robert. Privacy amplification by public discussion. *SIAM Journal on Computing*, 17(2):210–229, April 1988.

[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.

[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.

[CG14a] Mahdi Cheraghchi and Venkatesan Guruswami. Capacity of non-malleable codes. In *ITCS*, pages 155–168, 2014.

[CG14b] Mahdi Cheraghchi and Venkatesan Guruswami. Non-malleable coding against bit-wise and split-state tampering. In *TCC*, pages 440–464, 2014.

[CGL16] Eshan Chattopadhyay, Vipul Goyal, and Xin Li. Non-malleable extractors and codes, with their many tampered extensions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, 2016.

[CKOR10] N. Chandran, B. Kanukurthi, R. Ostrovsky, and L. Reyzin. Privacy amplification with asymptotically optimal entropy loss. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 785–794, 2010.

[CL16] Eshan Chattopadhyay and Xin Li. Explicit non-malleable extractors, multi-source extractors and almost optimal privacy amplification protocols. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, 2016.

[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.

[Coh16a] Gil Cohen. Making the most of advice: New correlation breakers and their applications. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, 2016.

[Coh16b] Gil Cohen. Non-malleable extractors - new tools and improved constructions. In *Proceedings of the 31st Annual IEEE Conference on Computational Complexity*, 2016.

[Coh16c] Gil Cohen. Non-malleable extractors with logarithmic seeds. Technical Report TR16-030, ECCC, 2016.

[Coh16d] Gil Cohen. Two-source extractors for quasi-logarithmic min-entropy and improved privacy amplification protocols. Technical Report TR16-114, ECCC: Electronic Colloquium on Computational Complexity, 2016.

[CRS14] Gil Cohen, Ran Raz, and Gil Segev. Non-malleable extractors with short seeds and applications to privacy amplification. *SIAM Journal on Computing*, 43(2):450–476, 2014.

[CS16] Gil Cohen and Leonard Schulman. Extractors for near logarithmic min-entropy. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, 2016.

[CZ14] Eshan Chattopadhyay and David Zuckerman. Non-malleable codes against constant split-state tampering. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, pages 306–315, 2014.

[CZ16] Eshan Chattopadhyay and David Zuckerman. Explicit two-source extractors and resilient functions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, 2016.

[DKO13] Stefan Dziembowski, Tomasz Kazana, and Maciej Obremski. Non-malleable codes from two-source extractors. In *CRYPTO (2)*, pages 239–257, 2013.

[DKRS06] Y. Dodis, J. Katz, L. Reyzin, and A. Smith. Robust fuzzy extractors and authenticated key agreement from close secrets. In *Advances in Cryptology — CRYPTO '06, 26th Annual International Cryptology Conference, Proceedings*, pages 232–250, 2006.

[DKSS09] Zeev Dvir, Swastik Kopparty, Shubhangi Saraf, and Madhu Sudan. Extensions to the method of multiplicities, with applications to kakeya sets and mergers. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, 2009.

[DLWZ14] Yevgeniy Dodis, Xin Li, Trevor D. Wooley, and David Zuckerman. Privacy amplification and non-malleable extractors via character sums. *SIAM Journal on Computing*, 43(2):800–830, 2014.

[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.

[DPW10] Stefan Dziembowski, Krzysztof Pietrzak, and Daniel Wichs. Non-malleable codes. In *ICS*, pages 434–452, 2010.

[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.

[DY13] Yevgeniy Dodis and Yu Yu. Overcoming weak expectations. In *10th Theory of Cryptography Conference*, 2013.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. *Journal of the ACM*, 56(4), 2009.

[KR09] B. Kanukurthi and L. Reyzin. Key agreement from close secrets over unsecured channels. In *EUROCRYPT 2009, 28th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2009.

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

[Li12a] Xin Li. Design extractors, non-malleable condensers and privacy amplification. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 837–854, 2012.

[Li12b] Xin Li. Non-malleable extractors, two-source extractors and privacy amplification. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 688–697, 2012.

[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. Non-malleable condensers for arbitrary min-entropy, and almost optimal protocols for privacy amplification. In *12th IACR Theory of Cryptography Conference*, pages 502–531. Springer-Verlag, 2015. LNCS 9014.

[Li15b] Xin Li. Three source extractors for polylogarithmic min-entropy. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, 2015.

[Li16] Xin Li. Improved two-source extractors, and affine extractors for polylogarithmic entropy. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, 2016.

[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.

[Mek15] Raghu Meka. Explicit resilient functions matching Ajtai-Linial. *CoRR*, abs/1509.00092, 2015.

[MW97] Ueli M. Maurer and Stefan Wolf. Privacy amplification secure against active adversaries. In *Advances in Cryptology — CRYPTO '97, 17th Annual International Cryptology Conference, Proceedings*, 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.

[RW03] Renato Renner and Stefan Wolf. Unconditional authenticity and privacy from an arbitrarily weak secret. In *Advances in Cryptology — CRYPTO '03, 23rd Annual International Cryptology Conference, Proceedings*, pages 78–95, 2003.

[Vad04] Salil P. Vadhan. Constructing locally computable extractors and cryptosystems in the bounded-storage model. *J. Cryptology*, 17(1):43–77, 2004.

# A The error in [ADKO15]

The error in [ADKO15] appears in the proof of Theorem 26 (Section 5.3), which reduces two look-ahead tampering to a $t$-split tampering. Specifically, to prove Equation (9) there one needs to argue about the quantity $H_\infty(L_i|\mathsf{Var}_i) = H_\infty(L_i|Z_1, \cdots, Z_{i-1})$. The claim is that $H_\infty(L_i|Z_1, \cdots, Z_{i-1}) \geq n/2$ because $L_i$ is a uniform string on $n$ bits, and the size of $(Z_1, \cdots, Z_{i-1})$ is at most $n/2$. However, this is not true. The only thing one can make sure is that the size of $(h_1(U^{(1)}, Z_1), \cdots, h_{i-1}(U^{(i-1)}, Z_{i-1}))$ is at most $n/2$, as written in the proof. But these are functions of $(Z_1, \cdots, Z_{i-1})$ and only output partial information. By examining the definition of $\{Z_i\}$, one can see that each $Z_i$ has $m \cdot 2^m$ bits, thus the size of $(Z_1, \cdots, Z_{i-1})$ can be up to $tm2^m$. Therefore, in order to make sure this is less than $n/2$, one needs $n \geq 2tm2^m$ in the theorem, rather than $n \geq 2tm$ as currently written.

We note that at this time, it is still not clear whether the proof can be fixed.



ECCC ISSN 1433-8092

http://eccc.hpi-web.de