Non-Malleable Extractors with Shorter Seeds and Privacy Amplification

Yanqing Yao[a,b], Zhoujun Li[a,c]

[a]State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[b]Department of Computer Science, New York University, New York 10012, USA
[c]Beijing Key Laboratory of Network Technology, Beihang University, Beijing, China
yaoyanqing1984@sina.com,lizj@buaa.edu.cn

Abstract. Motivated by the problem of how to communicate over a public channel with an active adversary, Dodis and Wichs [DW09] introduced the notion of a non-malleable extractor. A non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ takes two inputs, a weakly-random W and a uniformly random seed S, and outputs a string which is nearly uniform, given S as well as $\mathsf{nmExt}(W, \mathcal{A}(S))$, for an arbitrary function $\mathcal{A}$ with $\mathcal{A}(S) \neq S$.

In this paper, we improve the error estimation of Raz's extractor, which plays an extremely important role in the constraints of the non-malleable extractor parameters including the seed length. Then we present an improved explicit construction of non-malleable extractors, where the seed length is shorter than that by Cohen, Raz and Segev [CCC12]. More precisely, we construct an explicit $(1016, \frac{1}{2}) - 1-$non-malleable extractor $\mathsf{nmExt} : \{0,1\}^{2^{10}} \times \{0,1\}^d \to \{0,1\}$ with seed length 19, while it should be no less than $\frac{46}{63} + 66$ according to Cohen et al. in CCC'12. Therefore, it beats the condition "$2.01 \cdot \log n \leq d \leq n$" in CCC'12, since d is just $1.9 \cdot \log n$ in our construction. We also give a general explicit construction of non-malleable extractors and analyze the simplification of the constraints on the parameters. Finally, we give their application to privacy amplification.

Keywords: extractors; non-malleable; seed length; privacy amplification

1 Introduction

Randomness extractors are functions that convert weakly random sources into nearly uniform bits. Though the motivation of extractors is to simulate randomized algorithms with weak random sources as might arise in nature, randomness extractors have been successfully applied to coding theory, cryptography, complexity, etc. [9, 11, 19]. In this paper, we focus on the extractors that are applied to privacy amplification. In this scenario, two parties Alice and Bob share a weakly random secret key $W \in \{0,1\}^n$. The secret source W may be a human-memorizable password, some biometric data, and physical sources, which are themselves weakly random, or a uniform secret which may have been partially leaked to an adversary Eve. Thus, only the min-entropy of W is guaranteed. Alice and Bob interact over a public communication channel in order to securely agree on a nearly uniform secret key $R \in \{0,1\}^m$ in the presence of the adversary, Eve, who can see every message transmitted in the public channel. The min-entropy of W and the length of the public seed are two main measures of efficiency in this setting. If Eve is passive, a (strong) randomness extractor yields the following solution: Alice sends a uniformly random seed S to Bob, then they both compute $R = \mathsf{Ext}(W, X)$ as the nearly uniform secret key [15]. If Eve is active (i.e., it may change the messages in arbitrary ways), some protocols have been proposed to achieve this goal [4–8, 10–12, 18, 20].

As a major progress, Dodis and Wichs [8] introduced non-malleable extractors to study privacy amplification protocols, where the attacker is active and computationally unbounded. If an attacker sees a random seed S and modifies it into an arbitrarily related seed S', then the relationship between $R = \mathsf{Ext}(W, S)$ and $R' = \mathsf{Ext}(W, S')$ is bounded to avoid related key attacks. More formally, a non-malleable extractor is a function $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ that takes two inputs, a weakly-random secret source W with min-entropy k and uniformly random seed S, and outputs a string which is $\epsilon-$close to uniform in statistical distance, given S as well as $\mathsf{nmExt}(W, \mathcal{A}(S))$, for an arbitrary function $\mathcal{A}$ with $\mathcal{A}(S) \neq S$. They proved that $(k, \epsilon)-$non-malleable extractors exist as long as $k > 2m + 3\log\frac{1}{\epsilon} + \log d + 9$ and $d > \log(n-k+1) + 2\log\frac{1}{\epsilon} + 7$.

The first explicit non-malleable extractor was constructed by Dodis, Li, Wooley and Zuckerman [7]. It works for any weakly random input source with the min-entropy $k > \frac{n}{2}$ and uniformly random seed of length $d = n$ (It works even if the seed has entropy only $\Theta(m + \log n)$). However, when outputting more than a logarithmic number of bits, its efficiency relies on a longstanding conjecture on the distribution of prime numbers. Based on the inner product function, Li [11] later showed the first construction of non-malleable extractors for any weakly random source with the min-entropy $k = (\frac{1}{2} - \delta) \cdot n$ for some constant $\delta > 0$, and a conditional (semi-explicit) construction that can potentially achieve $k = \alpha n$ for any constant $\delta > 0$.

In 2012, an alternative explicit construction based on the extractor of Raz [17] was given by Cohen, Raz, and Segev in [5]. Without using any conjecture, their construction works for any weakly random source with the min-entropy $k = (\frac{1}{2} + \delta) \cdot n$ and uniformly random seed of length $d \geq \frac{23}{\delta} \cdot tm + 2\log n$ (see Theorem 1 for details). However, their result suffers from some drawbacks: The non-malleable extractor is constructed based on the explicit seeded extractor of Raz [17], while the error estimation in that construction is too rough. Moreover, though one main purpose of Cohen et al. in [5] is to shorten the length of the seed, the lower bound of the seed length is not optimal.

OUR CONTRIBUTIONS AND TECHNIQUES.

• We improve the error estimation of Raz's extractor in SOTC'05 [17], a special case of which was used by Cohen et al. in CCC'12 [5]. For simplicity, denote γ_1 as the error of the extractor in [5], and γ_2 as the counterpart in this paper. Recall that $\gamma_1 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot (2\epsilon)^{\frac{1}{k}}$ in [5] under the assumption that $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k$ and $0 < \delta < \frac{1}{2}$. If $\epsilon \geq \frac{1}{2^{(\frac{1}{2}-\delta)n+1}}$, then $\gamma_1 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot (2\epsilon)^{\frac{1}{k}} \geq 1$. In this case, the error estimation is meaningless. One main reason is that in those proofs, the partition method about the sum [5, 17] which bounds the error didn't capture the essence of the biased sequence for linear tests. In this paper, we propose another partition method and give a better bound on the sum by employing the combination and permutation formulas. Correspondingly, the error is $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot \{2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 + [1 - 2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1] \cdot \epsilon\}^{\frac{1}{k}}$. Since $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k$ and $2^{-\frac{dk}{2}} \cdot k^k > 2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1$ for any even integer k, we get $\gamma_1 > \gamma_2$. To simplify this bound, let k be a specific value. For instance, let $k = 4$, then the error $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)n}{4}} \cdot [2^{-2d} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$.

• Based on the improvement of the error estimation, we present an explicit construction of non-malleable extractors, which is an improvement of the construction of Cohen et al. in CCC'12 [5] in the sense that the seed length can be shorter. More concretely, we present an explicit $(1016, \frac{1}{2}) - 1-$non-malleable extractor $\mathsf{nmExt} : \{0,1\}^{2^{10}} \times \{0,1\}^d \rightarrow \{0,1\}$ with seed length $d = 19$, which beats the condition "$2.01 \cdot \log n \leq d \leq n$" in [5], since d is just $1.9 \cdot \log n$ in our construction while it should be no less than $\frac{46}{63} + 66$ according to [5]. We also give a general explicit construction of non-malleable extractors and analyze the simplification of the constraints on the parameters. Similar to [5], the non-malleable extractors are also constructed by using biased variable sequence for linear tests. In the proof, the error of the extractor impacts greatly on the constraints of the parameters including the length of the seed, the min-entropy of the source, the error of the non-malleable extractor.

ORGANIZATION. The remainder of the paper is organized as follows. In Section 2, we review some notations, concepts, and results. In section 3, we show an existing central lemma about the error estimation of Raz's Extractor and improve it by proposing a new partition method. In section 4, we propose the explicit construction of non-malleable extractors with shorter seed length compared with that in [5]. Then we show a general explicit construction of non-malleable extractors and analyze the simplification of the constraints on the parameters. In Section 5, we show how the non-malleable extractors can be applied to privacy amplification. Section 6 concludes the paper.

2 Preliminaries

In this section, we'll recall some notations, definitions, and results that will be used later.

For any positive integer n, denote $[n] = \{1, 2, \ldots, n\}$. Denote U_m as the uniformly random distribution over $\{0,1\}^m$. We measure the distance between two distributions by the $\mathcal{L}_1$ norm in order to be consistent with [5]. We say the distribution X is ϵ-close to the distribution Y if $\|X - Y\|_1 = \sum_s |Pr[X = s] - Pr[Y =$

$s]| \le \epsilon$ [1]. The statistical distance of X and Y is defined as $\mathsf{SD}(X,Y) = \frac{1}{2}\|X - Y\|_1$. It's well known that for any function f, $\mathsf{SD}(f(X), f(Y)) \le \mathsf{SD}(X,Y)$. Denote $\mathsf{SD}((X_1,X_2),(Y_1,Y_2)|Z)$ as the abbreviation of $\mathsf{SD}((X_1,X_2,Z),(Y_1,Y_2,Z))$.

The *min-entropy* of a random variable W is $H_\infty(W) = -\log \max_w Pr(W = w)$. For $W \in \{0,1\}^n$, we say W is an *(n,k)-source* if $H_\infty(W) \ge k$, and we call W has *entropy rate* $\frac{H_\infty(W)}{n}$. We say that a source (i.e., a random variable) is a *weak source* if its distribution is not uniform. We say W is a *flat source* if it is a uniform distribution over some subset $S \subseteq \{0,1\}^n$. Chor and Goldreich [3] observed that the distribution of any (n,k)-source is a convex combination of distributions of flat (n,b)-sources. Therefore, for general weak sources, it will be enough to consider flat sources instead in most cases.

Definition 1. A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a $(k,\epsilon)-$*seeded extractor* if for every (n,k)-source W and an independent uniformly random variable S (called *seed*) over $\{0,1\}^d$, the distribution of $\mathsf{Ext}(W,S)$ is ϵ-close to U_m. ϵ is called the *error* of the extractor. A seeded extractor is a *strong* $(k,\epsilon)-$*extractor* if for W and S as above, the distribution of $(\mathsf{Ext}(W,S),S)$ is ϵ-close to (U_m,U_d).

Definition 2. A random variable Z over $\{0,1\}$ is *ϵ-biased* if $bias(Z) = |Pr[Z = 0] - Pr[Z = 1]| \le \epsilon$ (i.e., Z is ϵ-close to uniform). A sequence of 0-1 random variables $Z_1,\ldots,Z_N$ is *ϵ-biased for linear tests of size k* if for any nonempty $\tau \subseteq [N]$ with $|\tau| \le k$, the random variable $Z_\tau = \oplus_{i\in\tau} Z_i$ is $\epsilon-$biased. We also say that the sequence $Z_1, Z_2, \ldots, Z_N$ *$\epsilon-$fools linear tests of size k.*

For every k', $N \ge 2$, variables $Z_1,\cdots,Z_N$ as above can be explicitly constructed using $2 \cdot \lceil \log(1/\epsilon) + \log k' + \log\log N \rceil$ random bits [1].

The Extractor of Raz. Raz [17] constructed an extractor based on a sequence of 0-1 random variables that have small bias for linear tests of a certain size. Let $Z_1,\cdots,Z_{m\cdot 2^d}$ be 0-1 random variables that are ϵ-biased for linear tests of size k' that are constructed using n random bits. The set of indices $[m\cdot 2^d]$ can be considered as the set $\{(i,s) : i \in [m], s \in \{0,1\}^d\}$. Define $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ by $\mathsf{Ext}(w,s) = Z_{(1,s)}(w)||Z_{(2,s)}(w)\ldots||Z_{(m,s)}(w)$, where "$||$" is the concatenation operator. Raz proposed that Ext is a seeded extractor with good parameters [17].

Cohen et al. [5] proved that the above extractor is in fact non-malleable. We'll also construct non-malleable extractors based on it. The formal definition of non-malleable extractors is as follows.

Definition 3. (see [8]) A function $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ is an *adversarial function* if for every $s \in \{0,1\}^d$ it holds that $\mathcal{A}(s) \ne s$. A function $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a *(k,ϵ)-non-malleable extractor* if for every (n,k)-source W, independent uniformly random variable S over $\{0,1\}^d$, and every

[1] In other literature (e.g., [11, 21]), X is ϵ-close to Y if $\frac{1}{2}\|X - Y\|_1 = \frac{1}{2}\sum_s |Pr[X = s] - Pr[Y = s]| \le \epsilon$.

adversarial function $\mathcal{A} : \{0,1\}^d \rightarrow \{0,1\}^d$,

$$\|(\mathsf{nmExt}(W, S), \mathsf{nmExt}(W, \mathcal{A}(S)), S) - (U_m, \mathsf{nmExt}(W, \mathcal{A}(S)), S)\|_1 \leq \epsilon.$$

ϵ is called the *error* of the extractor.

Definition 4. (see [5]) For any positive integer t, define $\mathcal{B} : \{0,1\}^d \rightarrow \{0,1\}^{td}$ as $\mathcal{B}(s) = \mathcal{A}_1(s)||\mathcal{A}_2(s)\cdots||\mathcal{A}_t(s)$, where for all $i \in [t]$, $\mathcal{A}_i$ is an adversarial function $\mathcal{A}_i : \{0,1\}^d \rightarrow \{0,1\}^d$ and "$||$" is the concatenation operator. We say $\mathcal{B}$ is a *$t-$adversarial function.* A function $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m$ is a $(k, \epsilon) -$ *t-non-malleable extractor* if for every (n, k)-source W, independent uniformly random variable S over $\{0,1\}^d$, and every t-adversarial function $\mathcal{B} : \{0,1\}^d \rightarrow \{0,1\}^{td}$,

$$\begin{aligned}&\|(\mathsf{nmExt}(W, S), \mathsf{nmExt}(W, \mathcal{A}_1(S)), \ldots, \mathsf{nmExt}(W, \mathcal{A}_t(S)), S)\\ &- (U_m, \mathsf{nmExt}(W, \mathcal{A}_1(S)), \ldots, \mathsf{nmExt}(W, \mathcal{A}_t(S)), S)\|_1 \leq \epsilon.\end{aligned}$$

One-time message authentication code (MAC) is used to guarantee that the received message is sent by a specified legitimate sender in an unauthenticated channel. Formally,

Definition 5. A family of functions $\{\mathsf{MAC}_r : \{0,1\}^v \rightarrow \{0,1\}^\tau\}_{r\in\{0,1\}^m}$ is a *ε-secure (one-time) message authentication code* (MAC) if for any μ and any function $f : \{0,1\}^\tau \rightarrow \{0,1\}^v \times \{0,1\}^\tau$, it holds that,

$$Pr_{r\leftarrow U_m}[\mathsf{MAC}_r(\mu') = \sigma' \wedge \mu' \neq \mu|(\mu', \sigma') = f(\mathsf{MAC}_r(\mu))] \leq \varepsilon.$$

The main theorem about the explicit construction of non-malleable extractors proposed in [5] is as follows.

Theorem 1. (see [5]) *For any integers n, d, m and t, and for any $0 < \delta < \frac{1}{2}$ such that*

$$d \geq \frac{23}{\delta} \cdot tm + 2\log n,$$

$$n \geq \frac{160}{\delta} \cdot tm,$$

$$\delta \geq 10 \cdot \frac{\log(nd)}{n},$$

there exists an explicit $((\frac{1}{2} + \delta) \cdot n, 2^{-m}) -$ t-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m$.

Proof. See [5]. Omitted.

For simplicity, we only consider $t = 1$ in the rest of the paper.

3 The Error Estimation and its Improvement of Raz's Extractor

In this section, we first recall the central lemma used in [5], which is a special case about the error estimation of Raz's Extractor [17]. Then we point out the flaw in the proof and improve its error estimation. Afterwards, we compare our result with the original one and roughly show the role of the improvement.

3.1 A Special Case of Raz's Extractor

The central lemma used in [5] is as follows, the proof of which is essentially the same as that in [17]. It can be considered as a special case of Raz's Extractor [17].

Lemma 1.(see Lemma 4.1 in [5]) *Let $D = 2^d$. Let $Z_1, \cdots, Z_D$ be 0-1 random variables that are ϵ-biased for linear tests of size k' that are constructed using n random bits. Define $\mathsf{Ext}^{(1)}$: $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ by $\mathsf{Ext}^{(1)}(w,s) = Z_s(w)$, that is, $\mathsf{Ext}^{(1)}(w,s)$ is the random variable Z_s, when using w as the value of the n bits needed to produce $Z_1, \cdots, Z_D$. Then, for any $0 < \delta < \frac{1}{2}$ and even integer $k = k'$ such that $k \cdot (\frac{1}{\epsilon})^{\frac{1}{k}} \leq D^{\frac{1}{2}}$, the function $\mathsf{Ext}^{(1)}$ is a $((\frac{1}{2}+\delta)\cdot n, \gamma)$-seeded-extractor, with $\gamma = (\epsilon \cdot 2^{(\frac{1}{2}-\delta)n+1})^{\frac{1}{k}}$.*

Recall that the proof sketch is as follows.

Proof. Let W be a $(n, (\frac{1}{2}+\delta)\cdot n)$-source. Let S be a random variable that is uniformly distributed over $\{0,1\}^d$ and is independent of W. We will show that the distribution of $\mathsf{Ext}^{(1)}(W,S)$ is $\gamma-$close to uniform. As in [3], it is enough to consider the case where W is uniformly distributed over a set $W' \subseteq \{0,1\}^n$ of size $2^{(1/2+\delta)n}$. For every $w \in \{0,1\}^n$ and $s \in \{0,1\}^d$ denote

$$e(w,s) = (-1)^{Z_s(w)}.$$

Claim 1. *For any $r \in [k]$ and any different $s_1, \cdots, s_r \in \{0,1\}^d$,*

$$\sum_{w \in \{0,1\}^n} \prod_{j=1}^{r} e(w, s_j) \leq \epsilon \cdot 2^n.$$

Proof.

$$\begin{aligned} \sum_{w\in\{0,1\}^n} \prod_{j=1}^{r} e(w,s_j) &= \sum_{w\in\{0,1\}^n} \prod_{j=1}^{r} (-1)^{Z_{s_j}(w)} \\ &= \sum_{w\in\{0,1\}^n} (-1)^{Z_{s_1}(w)\oplus\cdots\oplus Z_{s_r}(w)}, \end{aligned}$$

and since $Z_{s_1}(w) \oplus \cdots \oplus Z_{s_r}(w)$ is $\epsilon-$biased, the last sum is at most $\epsilon \cdot 2^n$. □

The $\mathcal{L}_1$ distance of $\mathsf{Ext}^{(1)}(W,S)$ and U is

$$\begin{aligned} &\|\mathsf{Ext}^{(1)}(W,S) - U\|_1 \\ =& |Pr[\mathsf{Ext}^{(1)}(W,S)=0] - Pr[U=0]| + |Pr[\mathsf{Ext}^{(1)}(W,S)=1] - Pr[U=1]| \\ =& |Pr[\mathsf{Ext}^{(1)}(W,S)=0] - Pr[\mathsf{Ext}^{(1)}(W,S)=1]| \\ =& |\frac{1}{2^{(\frac{1}{2}+\delta)n}} \cdot \frac{1}{2^d}(\sum_{w\in W'} \sum_{s\in\{0,1\}^d} e(w,s))|. \end{aligned}$$

Denote $\gamma(W,S)=\frac{1}{2^{(\frac{1}{2}+\delta)n}}\cdot\frac{1}{2^d}(\sum_{w\in W'}\sum_{s\in\{0,1\}^d}e(w,s))$.

Define $f:[-1,1]\to[-1,1]$ by $f(z)=z^k$, then f is a convex function for any even positive integer k.

Thus, by a convexity argument, we have

$$\begin{aligned}
&2^{(\frac{1}{2}+\delta)n}\cdot(2^d\cdot\gamma(W,S))^k\\
&=2^{(\frac{1}{2}+\delta)n}\cdot\{\sum_{w\in W'}[\frac{1}{2^{(1/2+\delta)n}}\sum_{s\in\{0,1\}^d}e(w,s)]\}^k\\
&\leq 2^{(\frac{1}{2}+\delta)n}\cdot\{\sum_{w\in W'}\frac{1}{2^{(1/2+\delta)n}}[\sum_{s\in\{0,1\}^d}e(w,s)]^k\}\\
&\leq\sum_{w\in\{0,1\}^n}[\sum_{s\in\{0,1\}^d}e(w,s)]^k\\
&=\sum_{w\in\{0,1\}^n}\sum_{s_1,\cdots,s_k\in\{0,1\}^d}\prod_{j=1}^{k}e(w,s_j)\\
&=\sum_{s_1,\cdots,s_k\in\{0,1\}^d}\sum_{w\in\{0,1\}^n}\prod_{j=1}^{k}e(w,s_j).
\end{aligned}$$

The sum over $s_1,\cdots,s_k\in\{0,1\}^d$ is broken into two sums. The first sum is over $s_1,\cdots,s_k\in\{0,1\}^d$ such that in each summand, at least one s_j is different than all other elements in $\{s_1,\cdots,s_k\}$, and the second sum is over $s_1,\cdots,s_k\in\{0,1\}^d$ such that in each summand every s_j is identical to at least one other element in $\{s_1,\cdots,s_k\}$. The number of summands in the first sum is trivially bounded by $2^{d\cdot k}$, and by Claim 1 each summand is bounded by $2^n\cdot\epsilon$. The number of summands in the second sum is bounded by $2^{d\cdot\frac{k}{2}}\cdot(\frac{k}{2})^k$, and each summand is trivially bounded by 2^n. Therefore,

$$\begin{aligned}
2^{(\frac{1}{2}+\delta)n}\cdot 2^{d\cdot k}\cdot\gamma(W,S)^k&\leq 2^n\cdot\epsilon\cdot 2^{d\cdot k}+2^n\cdot 2^{d\cdot\frac{k}{2}}\cdot(\frac{k}{2})^k\\
&\leq 2\cdot 2^n\cdot\epsilon\cdot 2^{d\cdot k},
\end{aligned}$$

where the last inequality follows by the assumption that $k\cdot(1/\epsilon)^{1/k}\leq D^{\frac{1}{2}}$. That is,

$$|\gamma(W,S)|\leq(\epsilon\cdot 2^{(\frac{1}{2}-\delta)n+1})^{\frac{1}{k}}.$$

□

The above partition method about the sum over $s_1,\cdots,s_k\in\{0,1\}^d$ is not optimal, since it doesn't capture the essence of random variable sequence that is biased for linear tests. Moreover, the bounds of the number of summands in the two sums are too large. The same problem exists in [17].

In fact, when every s_j is identical to at least one other element in $\{s_1, \cdots, s_k\}$ under the assumption that at least one s_j appears odd times in $\{s_1, \cdots, s_k\}$, the summand $\sum\limits_{w\in\{0,1\}^n} \prod\limits_{j=1}^{k} e(w, s_j)$ is still upper bounded by $2^n \cdot \epsilon$, since $\sum\limits_{w\in\{0,1\}^n} \prod\limits_{j=1}^{k} e(w, s_j) = \sum\limits_{w\in\{0,1\}^n} \prod\limits_{j=1}^{k} (-1)^{Z_{s_j}(w)} = \sum\limits_{w\in\{0,1\}^n} (-1)^{Z_{s_1}(w)\oplus\cdots\oplus Z_{s_k}(w)}$ and $Z_1, \cdots, Z_D$ are 0-1 random variables that are ϵ-biased for linear tests of size k'. However, in this case the upper bound of the summand $\sum\limits_{w\in\{0,1\}^n} \prod\limits_{j=1}^{k} e(w, s_j)$ was considered to be 2^n in [5, 17].

3.2 Improvement of the Error Estimation

We improve the error estimation of Raz's extractor as follows. Unlike [5, 17], we present another partition method of the sum in the following proof. The combination and permutation formulas are exploited to show a tight bound of the sum. Correspondingly, the error can be reduced.

Theorem 2. *Let $D = 2^d$. Let $Z_1, \cdots, Z_D$ be 0-1 random variables that are ϵ-biased for linear tests of size k' that are constructed using n random bits. Define $\mathsf{Ext}^{(1)}$: $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ by $\mathsf{Ext}^{(1)}(w, s) = Z_s(w)$, that is, $\mathsf{Ext}^{(1)}(w, s)$ is the random variable Z_s, when using w as the value of the n bits needed to produce $Z_1, \cdots, Z_D$. Then, for any even integer $k = k'$, the function $\mathsf{Ext}^{(1)}$ is a (α, γ)-seeded-extractor, with*

$$\gamma = 2^{\frac{n-\alpha}{k}} \cdot [2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}.$$

Proof. We improve the proof by proposing another method for partitioning the sum $\sum\limits_{s_1,\cdots,s_k\in\{0,1\}^d} \sum\limits_{w\in\{0,1\}^n} \prod\limits_{j=1}^{k} e(w, s_j)$ into two sums. The first sum is over $s_1, \cdots, s_k \in \{0,1\}^d$ such that in each summand, at least one s_j appears odd times in $\{s_1, \cdots, s_k\}$, and the second sum is over $s_1, \cdots, s_k \in \{0,1\}^d$ such that in each summand every s_j appears even times in $\{s_1, \cdots, s_k\}$. Then the number of summands in the second sum is

$$\frac{C_2^k \cdot C_2^{k-2} \cdot \cdots \cdot C_2^2}{P_{\frac{k}{2}}^{\frac{k}{2}}} \cdot 2^{\frac{dk}{2}} = \frac{k! \cdot \frac{1}{2^{\frac{k}{2}}}}{(\frac{k}{2})!} \cdot 2^{\frac{dk}{2}} = \frac{k!}{(\frac{k}{2})! \cdot 2^{\frac{k}{2}}} \cdot 2^{\frac{dk}{2}} = 2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1,$$

and each summand is 2^n. Therefore, the number of summands in the first sum is

$$2^{dk} - 2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1,$$

and by Claim 1 each summand is bounded by $2^n \cdot \epsilon$. Hence,

$$\begin{aligned}&2^\alpha \cdot 2^{d \cdot k} \cdot \gamma(W,S)^k \\ \leq\ & 2^n \cdot [2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1] + 2^n \cdot \epsilon \cdot [2^{dk} - 2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1]\end{aligned}$$

Therefore,

$$\begin{aligned}\gamma(W,S)^k &\leq \frac{2^n \cdot 2^{dk}}{2^\alpha \cdot 2^{d \cdot k}} \cdot [2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon] \\ &= 2^{n-\alpha} \cdot [2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]\end{aligned}$$

Consequently, we have

$$|\gamma(W,S)| \leq 2^{\frac{n-\alpha}{k}} \cdot [2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}.$$

□

3.3 Comparison

For simplicity, in the rest of the paper, denote γ_1 as the error of the extractor in Lemma 1, and γ_2 as the counterpart in Theorem 2.

Proposition 1. $2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \leq 2^{d \cdot \frac{k}{2}} \cdot (\frac{k}{2})^k$ *for any positive even integer* k*, and "=" holds iff* $k = 2$*.*

Proof. When $k = 2$, it's trivial that $2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 = 2^{d \cdot \frac{k}{2}} \cdot (\frac{k}{2})^k$. In the following, we only consider any positive even integer k with $k > 2$.

Since $\frac{k!}{(\frac{k}{2})!} < \frac{k^k}{2^{\frac{k}{2}}}$, we have $\frac{k!}{(\frac{k}{2})! \cdot 2^{\frac{k}{2}}} < \frac{k^k}{2^k}$. Hence,

$$2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 = 2^{\frac{dk}{2}} \cdot \frac{k!}{(\frac{k}{2})! \cdot 2^{\frac{k}{2}}} < 2^{\frac{dk}{2}} \cdot \frac{k^k}{2^k}.$$

□

- When k is large enough, $(\frac{k}{2})^k$ is much greater than $(k-1) \cdot (k-3) \cdot \cdots \cdot 1$. For instance, when $k = 6$, we have $(\frac{k}{2})^k = 729$ and $(k-1) \cdot (k-3) \cdot \cdots \cdot 1 = 15$. Therefore, "The number of summands in the second sum is $2^{\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1$, and each summand is 2^n." in the proof of Theorem 2 is a great improvement on "The number of summands in the second sum is bounded by $2^{d \cdot \frac{k}{2}} \cdot (\frac{k}{2})^k$, and each summand is trivially bounded by 2^n." in the proof of Lemma 1.

- The error estimation of the extractor in Theorem 1 is better than that in Lemma 1. $\gamma_1 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot (2\epsilon)^{\frac{1}{k}}$ in [5] under the assumption that $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k$ and $0 < \delta < \frac{1}{2}$, and $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot [2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}} = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot \{2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 + [1 - 2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1] \cdot \epsilon\}^{\frac{1}{k}}$.

Since $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k$ and $2^{-\frac{dk}{2}} \cdot k^k > 2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \dots \cdot 1$ for any even integer k, we get $\gamma_1 > \gamma_2$. Moreover, according to the Stirling's Formula, $2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \dots \cdot 1 = 2^{-\frac{dk}{2}} \cdot \frac{k!}{(\frac{k}{2})! \cdot 2^{\frac{k}{2}}} \approx 2^{-\frac{dk}{2}} \cdot \frac{k^k}{(\frac{k}{2})^{\frac{k}{2}} \cdot 2^{\frac{k}{2}}} = 2^{-\frac{dk}{2}} \cdot \sqrt{k}^k$.

- If $\epsilon \geq \frac{1}{2^{(\frac{1}{2}-\delta)n+1}}$, then $\gamma_1 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot (2\epsilon)^{\frac{1}{k}} \geq 1$. In this case, the error estimation is meaningless.
- To simplify γ_2, let k be a specific value. For instance, let $k = 4$, then the error $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)n}{4}} \cdot [2^{-2d} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$.

3.4 Important Role

It should be noticed that γ constrains on the parameters in Theorem 1. The main idea of the proof about Theorem 1 is as follows. Assuming for contradiction that Ext is not a non-malleable extractor, then after some steps, an inequality $\gamma_1 > A$ is deduced, where A denotes a certain formula. On the other hand, from the assumption of Theorem 1, $\gamma_1 < A$ should hold. Thus Ext is a non-malleable extractor. Essentially, the constraints on the parameters in Theorem 1 are chosen according to the inequality $\gamma_1 < A$. From Proposition 1, we have $\gamma_1 > \gamma_2$ for any positive even integer $k \geq 4$. Therefore, we may relax the constraints on the parameters in Theorem 1 according to $\gamma_2 < A$. Correspondingly, the seed length may be further shortened.

4 Explicit Constructions of Non-malleable Extractors with Shorter Seed Length

In this section, we propose the explicit construction of a non-malleable extractor, which has shorter seed length than that in Theorem 1. Then we give a general explicit construction of non-malleable extractors, and analyze the simplification of the constraints on the parameters.

4.1 Constructions

We first review two lemmas that will be used later.

Lemma 2. (see [5]) *Let X be a random variable over $\{0,1\}^m$. Let Y, Z be two random variables. Then,*

$$\|(X,Y,Z) - (U_m,Y,Z)\|_1 = \mathbb{E}_{z \sim Z} \|(X,Y)|_{Z=z} - (U_m,Y)|_{Z=z}\|_1.$$

Lemma 3. (see [5]) *Let X be a random variable over $\{0,1\}^m$. Let Y be a random variable over $\{0,1\}^n$. Then*

$$\|(X,Y) - (U_m,Y)\|_1 \leq \sum_{\emptyset \neq \sigma \subseteq [m], \tau \subseteq [n]} bias(X_\sigma \oplus Y_\tau),$$

where X_i is the ith bit of X, Y_j is the jth bit of Y, $X_\sigma = \oplus_{i\in\sigma} X_i$, and $Y_\tau = \oplus_{j\in\tau} Y_j$.

Theorem 3. *Let $n = 2^{10}$, $d = 19$, $m = 1$, $\alpha = 1016$, $\xi = 2^\theta$, $\theta = -1$, and $t = 1$. Then there exists an explicit $(\alpha, \xi) - t$-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$.*

Proof Idea. We borrow the reductio ad absurdum approach in the proof of Theorem 1. The first part of the proof is almost the same as that in [5]. The second part jumps out of the idea in [5]. We exploit the error estimation of the extractor in Theorem 2 instead of Lemma 1. We use a trick such that the even integer k is just 4 instead of the largest even integer that is not larger than $\frac{\lceil \frac{\delta n}{8} \rceil}{(t+1)m}$. Therefore the extractor error can be simplified and complex assumptions to satisfy $k \cdot (\frac{1}{\epsilon})^{\frac{1}{k}} \leq (m \cdot 2^d)^{\frac{1}{2}}$ is omitted.

Proof. The explicit construction we present is the extractor constructed in [17]. Alon et al. [1] observed that for every k', $N \geq 2$, the sequence of 0-1 random variables $Z_1, \cdots, Z_N$ that is ϵ-biased for linear tests of size k' can be explicitly constructed using $2 \cdot \lceil \log(1/\epsilon) + \log k' + \log\log N \rceil$ random bits. Therefore, let $D = 2^d$ and $\epsilon = 2^{-\frac{n}{2}+r}$ with $r = 1 + \log k' + \log\log D$, then we can construct a sequence of 0-1 random variables $Z_1, \cdots, Z_D$ that is ϵ-biased for linear tests of size k' by using n random bits. Let $k' = 8$. Define $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ by $\mathsf{Ext}(w, s) = Z_s(w)$.

Let S be a random variable uniformly distributed over $\{0,1\}^d$.

Assume for contradiction that Ext is not a $(1016, \frac{1}{2})$-non-malleable-extractor. Then there exists a source W of length n with min-entropy α, and an adversarial-function $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ such that

$$\|(\mathsf{Ext}(W,S), \mathsf{Ext}(W,\mathcal{A}(S)), S) - (U_m, \mathsf{Ext}(W,\mathcal{A}(S)), S)\|_1 > \xi = 2^\theta.$$

As in [3], we assume that W is uniformly distributed over a set $W' \subseteq \{0,1\}^n$ of size α.

For every $s \in \{0,1\}^d$, let X_s be the random variable $\mathsf{Ext}(W, s)$. By Lemma 2 and 3, we have

$$\sum_{\emptyset \neq \sigma \subseteq [1], \tau \subseteq [1]} \mathbb{E}_{s\sim S}[bias((X_s)_\sigma \oplus (X_{\mathcal{A}(s)})_\tau)] > 2^\theta.$$

Let $\sigma^*, \tau^* \subseteq [1]$ be the indices of (one of) the largest summands in the above sum. For every $s \in \{0,1\}^d$, let

$$Y_s = (X_s)_{\sigma^*} \oplus (X_{\mathcal{A}(s)})_{\tau^*}.$$

There is a set $S'' \subseteq \{0,1\}^d$ satisfying that $|S''| > \frac{\xi \cdot 2^{d-2}}{2^{mt}(2^m-1)(t+1)^2} = \xi \cdot 2^{d-5}$. The S'' here is the same as that in the proof of Theorem 1 by replacing t there with 1 and the error 2^{-m} there with ξ. Please see [5] for details.

Define a random variable $Y_{S''}$ over $\{0,1\}$ as follows: To sample a bit from $Y_{S''}$, uniformly sample a string s from S'', and then independently sample a string

w uniformly from W'. The sampled value is $Y_s(w)$. We have that $bias(Y_{S''}) > \frac{\xi}{2^{mt+1}(2^m-1)(t+1)} = \frac{\xi}{2^3}$. For every $s \in S''$, let

$$Y'_s = Z_{(1,s)} \oplus (\oplus_{j \in \mathcal{T}^*} Z_{(j,\mathcal{A}(s))}),$$

where $Z_{(1,s)} = Z_s$.

Let $t = 1$ and $m = 1$ in Claim 7.2 of [5], we get the following claim.

Claim 2. *The set of random variables* $\{Y'_s\}_{s \in S''}$ $\epsilon-$*fools linear tests of size* $\frac{k'}{2}$.

We apply Theorem 2 on the random variables $\{Y'_s\}_{s \in S''}$. For simplicity of presentation we assume $|S''| = 2^{d'}$. By Theorem 2, the distribution of $\mathsf{Ext}^{(1)}(W, S'')$ is $\gamma-$biased for $\gamma = 2^{\frac{n-\alpha}{k}} \cdot [2^{-\frac{d'k}{2}} \cdot (k-1) \cdot (k-3) \cdot \dots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}$. Let $k = \frac{k'}{2} = 4$, then $\gamma = 2^{\frac{n-\alpha}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$. We note that $\mathsf{Ext}^{(1)}(W, S'')$ has the same distribution as $Y_{S''}$. In particular, both random variables have the same bias. Therefore, we get

$$2^{\frac{n-\alpha}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} \geq bias(Y_{S''}) > \frac{\xi}{2^3},$$

Moreover, since $2^{d'} = |S''| > \xi \cdot 2^{d-5}$, we have

$$2^{\frac{n-\alpha}{4}} \cdot [\xi^{-2} \cdot 2^{-2d+10} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} > 2^{\frac{n-\alpha}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} > \frac{\xi}{2^3}.$$

That is,

$$2^{-2d} > \frac{\xi^4 \cdot 2^{-12-n+\alpha} - \epsilon}{3(1-\epsilon) \cdot 2^{10} \cdot \xi^{-2}}, \qquad (a)$$

where $\epsilon = 2^{-\frac{n}{2}+r}$ and $r = 4 + \log d$.

On the other hand, from the assumption that $n = 2^{10}$, $\alpha = 1016$, $d = 19$, $m = 1$, $\xi = 2^{\theta}$, $\theta = -1$, we have

$$2^{-2d} < \frac{\xi^4 \cdot 2^{-12-n+\alpha} - \epsilon}{3(1-\epsilon) \cdot 2^{10} \cdot \xi^{-2}},$$

which is in contradiction to the inequality (a).

□

Comparison. In Theorem 1, the seed length d and the source length n should satisfy $d \geq \frac{23}{\delta}m + 2\log n$ with $0 < \delta < \frac{1}{2}$. However, in the above construction, we have $d = 1.9\log n$. In other words, let $n = 2^{10}$, $m = 1$, $\delta = \frac{504}{1024}$, and $t = 1$ in Theorem 1, then it can be easily verified that $n \geq \frac{160}{\delta} \cdot tm$. To construct an explicit $((\frac{1}{2}+\delta) \cdot n, 2^{-m}) - t$-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m$ (that is, an explicit $(1016, \frac{1}{2})$-1-non-malleable extractor nmExt), by Theorem 1, the seed length d should satisfy $d \geq \frac{23}{\delta} \cdot tm + 2\log n = \frac{46}{63} + 66$. Moreover, when $d \leq 2^{41}$, the precondition $\delta \geq 10 \cdot \frac{\log(nd)}{n}$ in Theorem 1 is satisfied. Meanwhile, by Theorem 3, the seed length d can just be 19. In this sense, our construction is much better than that of [5].

Theorem 4. *Assuming*

$$0 < 2^{\log 3-2\theta+4m+8}-2^{\log 3-\frac{n}{2}+4+\log d-2\theta+4m+8} \le 2^{2d+4\theta-8m-8-n+\alpha}-2^{2d-\frac{n}{2}+4+\log d}.$$

Then there exists an explicit $(\alpha, \xi = 2^{\theta}) - 1$*-non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$.

Proof. The proof is similar to Theorem 3. Please see the appendix for details.

Analysis of the Assumption. In order to construct an explicit non-malleable extractor, it's enough to guarantee that the parameters satisfies

$$0 < 2^{\log 3}\cdot(1-2^{-\frac{n}{2}+4+\log d})\cdot 2^{-2\theta+4m+8} \le 2^{2d+4\theta-8m-8-n+\alpha}-2^{2d-\frac{n}{2}+4+\log d}. \ (b)$$

For simplicity, denote

$$A' = \log 3 - 2\theta + 4m + 8,\ B' = \log 3 - \frac{n}{2} + 4 + \log d - 2\theta + 4m + 8,$$
$$C' = 2d + 4\theta - 8m - 8 - n + \alpha,\ D' = 2d - \frac{n}{2} + 4 + \log d,$$

then

$$(b)\ holds.\ \Leftrightarrow 0 < 2^{A'} - 2^{B'} \le 2^{C'} - 2^{D'}.$$

In the following, we discuss what happens under the assumption (b) in three cases.

Case 1. Assuming $A' \ge C'$ and $B' \ge D'$. Since "$B' \ge D'$" implies "$A' \ge C'$", we only need to consider $B' \ge D'$ (i.e., $\log 3 - 2\theta + 4m + 8 \ge 2d$). Let $1 - \epsilon = 1 - 2^{-\frac{n}{2}+4+\log d} = 2^{\rho'}$.

From $\log 3 + 8 + 4m \ge 2d + 2\theta$, $\alpha \le n$, $m \ge 1$, and $\theta < 0$, we get

$$\begin{aligned} -16 &> -8m - 8 + 4\theta - n + \alpha \\ &= (\log 3 + 8 + 4m) + 4\theta - 12m - 16 - \log 3 - n + \alpha \\ &\ge 2d + 2\theta + 4\theta - 12m - 16 - \log 3 - n + \alpha. \end{aligned}$$

Let $\rho' \ge -16$. Then we have $\rho' > 2d + 2\theta + 4\theta - 12m - 16 - \log 3 - n + \alpha$.

Therefore, $\log 3 + \rho' - 2\theta + 4m + 8 > 2d + 4\theta - 8m - 8 - n + \alpha$, which is in contradiction to the inequality (b).

Consequently, when $\epsilon \in (0, 1-2^{-16}]$, $A' \ge C'$, and $B' \ge D'$, (b) does not hold. From Theorem 2, only if ϵ is small enough, the corresponding seeded extractor is useful. Therefore, we assume that $\epsilon \in (0, 1 - 2^{-16}]$.

Case 2. Assuming $A' \ge C'$ and $B' < D'$, then it's in contradiction to the inequality (b).

Case 3. Assuming $A' < C'$, then it's trivial that $B' < D'$. Thus, we only need to consider $A' < C'$. Since $A' > B'$, we have $C' > D'$, that is, $4\theta - 8m - 12 - \frac{n}{2} + \alpha > \log d$.

Therefore, we obtain the following theorem.

Theorem 5. *Assuming* $\epsilon = 2^{-\frac{n}{2}+4+\log d} \in (0, 1-2^{-16}]$ *and*

$$2^{\log 3} \cdot (1 - 2^{-\frac{n}{2}+4+\log d}) \cdot 2^{-2\theta+4m+8} \leq 2^{2d+4\theta-8m-8-n+\alpha} - 2^{2d-\frac{n}{2}+4+\log d}.$$

Then there exists an explicit $(\alpha, \xi = 2^{\theta}) - 1$*-non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$.

In particular, the parameters of the non-malleable extractor can be chosen according to the inequality system

$$\begin{cases} \log 3 - 6\theta + 16 + 12m + n - \alpha < 2d \\ 4\theta - 8m - 12 - \frac{n}{2} + \alpha > \log d \\ 2^{-\frac{n}{2}+4+\log d} \leq 1 - 2^{-16} \end{cases} \quad (1)$$

then check whether they satisfy the inequality

$$2^{\log 3-2\theta+4m+8} - 2^{\log 3-\frac{n}{2}+4+\log d-2\theta+4m+8} \leq 2^{2d+4\theta-8m-8-n+\alpha} - 2^{2d-\frac{n}{2}+4+\log d}.$$

Remark. α can't be less than $\frac{n}{2}$, since $4\theta - 8m - 12 - \frac{n}{2} + \alpha > \log d$. We can also replace α with $(\frac{1}{2} + \delta)n$ where $0 < \delta < \frac{1}{2}$.

5 Application to Privacy Amplification

In this section, we show how the non-malleable extractor is applied to the privacy amplification protocol of Dodis and Wichs [7, 8] (also known as an information-theoretic key agreement protocol).

Roughly speaking, in this scenario, Alice and Bob share a shared weak secret W, the min-entropy of which is only guaranteed. They communicate over a public and unauthenticated channel to securely agree on a nearly uniform secret key R, where the attacker Eve is active and computationally unbounded. To achieve this goal, the protocol is designed as follows. Assuming we'll authenticate the seed S_0. Alice initiates the conversation by transmitting a uniformly random seed S to Bob. During this transmission, S may be modified by Eve into any value S'. Then Bob samples a uniform seed S_0, computes the authentication key $R' = \mathsf{nmExt}(W, S')$, and sends S_0 together with the authentication tag $T_0 = \mathsf{MAC}_{R'}(S_0)$ to Alice. At this point, Bob reaches the KeyDerived state and outputs $R_B = \mathsf{Ext}(W, S_0)$. During this transmission, (S_0, T_0) may be modified by Eve into any pair (S_0', T_0'). Alice computes the authentication key $R = \mathsf{nmExt}(W, S)$ and verifies that $T_0' = \mathsf{MAC}_R(S_0')$. If the verification fails then Alice rejects and outputs $R_A = \perp$. Otherwise, Alice reaches the KeyConfirmed state and outputs $R_A = \mathsf{nmExt}(W, S_0')$.

The security can be analyzed in two cases [5, 7]. Case 1: Eve does not modified the seed S in the first round. Then Alice and Bob share the same authentication key (i.e., $R' = R$), which is statistically close to a uniform distribution. Therefore, Eve has only a negligible probability of getting a valid authentication tag T_0' for any seed $S_0' \neq S_0$. Case 2: Eve does modify the seed S to a different seed S'.

Alice: W	Eve	Bob: W
Sample random S.		
	$S \longrightarrow S'$	
		Sample random S_0.
		$R' = \mathsf{nmExt}(W, S')$.
		$T_0 = \mathsf{MAC}_{R'}(S_0)$.
		Reach $\mathsf{KeyDerived}$ state.
		Output $R_B = \mathsf{Ext}(W, S_0)$.
	$(S_0', T_0') \longleftarrow (S_0, T_0)$	
$R = \mathsf{nmExt}(W, S)$.		
If $T_0' \neq \mathsf{MAC}_R(S_0')$, output $R_A = \perp$.		
Otherwise, reach $\mathsf{KeyConfirmed}$ state,		
and output $R_A = \mathsf{Ext}(W, S_0')$.		

Table 1. The Dodis-Wichs privacy amplification protocol.

Since T_0 is a deterministic function of S_0 and R', Eve may guess R'. According to the definition of non-malleable extractors, the authentication key R computed by Alice is still statistically close to a uniform distribution. Thus, again, the adversary has only a negligible probability of computing a valid authentication T_0' for any seed S_0' with respect to the authentication key R. Consequently, the above protocol is secure. The formal definition of privacy amplification protocol is given in the appendix.

Theorem 8. (see [5, 8]) *Assume that* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ *is a* $(k, \epsilon_{\mathsf{nmExt}})$*-non-malleable extractor,* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$ *is a strong* $(k - (d_1 + m_1) - \log \frac{1}{\epsilon'}, \epsilon_{\mathsf{Ext}})$*-extractor, and* $\{\mathsf{MAC}_r : \{0,1\}^{d_2} \to \{0,1\}^{\tau}\}_{r \in \{0,1\}^{m_1}}$ *is a* $\varepsilon_{\mathsf{MAC}}$*-secure message authentication code. Then for any integers* n *and* $k < n$*, the protocol in Table 1 is a 2-round* (n, k, m, δ)*-privacy amplification protocol, with communication complexity* $d_1 + d_2 + \tau$*, where* $\delta = \max\{\epsilon' + \epsilon_{\mathsf{Ext}}, \epsilon_{\mathsf{nmExt}} + \varepsilon_{\mathsf{MAC}}\}$.

The explicit non-malleable extractor in this paper can be applied to construct the above privacy amplification protocol with low communication complexity.

6 Conclusion

Non-malleable extractor is a powerful theoretical tool to study privacy amplification protocols, where the attacker is active and computationally unbounded.

In this paper, we improved the error estimation of Raz's extractor. Based on the improvement, we presented an improved explicit construction of non-malleable extractors with shorter seed length. More precisely, we constructed an explicit $(1016, \frac{1}{2})-1-$non-malleable extractor $\mathsf{nmExt} : \{0,1\}^{2^{10}} \times \{0,1\}^d \to \{0,1\}$ with seed length 19, while it should be no less than $\frac{46}{63} + 66$ according to Cohen et al. in CCC'12 [5]. We also gave a general explicit construction of non-malleable extractors and analyzed the simplification of the constraints on the parameters.

How to further simplify the constraints is an open problem. Finally, we showed their application to the privacy amplification protocol (or information-theoretic key agreement protocol).

Acknowledgments. We would like to thank Divesh Aggraval, Yevgeniy Dodis, Feng-Hao Liu, and Xin Li for helpful discussions. This work is supported by the Natural Science Foundation of China [grant numbers 60973105, 90718017, 61170189, 61170107 and 61300172]; the Fund for the Doctoral Program of Higher Education of China [grant numbers 20101303110004, 20111102130003 and 20121102120017]; the Fund of the State Key Laboratory of Software Development Environment [grant numbers SKLSDE-2011ZX-03 and SKLSDE-2012ZX-11]; the Fund of the Scholarship Award for Excellent Doctoral Student granted by Ministry of Education [grant number 400618]; and the Fund for CSC Scholarship Programme [grant number 201206020063].

References

1. N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple construction of almost k-wise independent random variables. Random Structures and Algorithms, 3(3): 289-304, 1992.
2. J. Bourgain. More on the sum-product phenomenon in prime fields and its applications. International Journal of Number Theory, 1: 1-32, 2005.
3. B. Chor and O. Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. SIAM Journal on Computing, 17(2): 230-261, 1988.
4. 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.
5. G. Cohen, R. Raz, and G. Segev. Non-malleable Extractors with Short Seeds and Applications to Privacy Amplification. IEEE Conference on Computational Complexity 2012: 298-308.
6. Y. Dodis, J. Katz, L. Reyzin, and A. Smith. Robust fuzzy extractors and authenticated key agreement from close secrets. In CRYPTO, pages 232-250, 2006.
7. Y. Dodis, X. Li, T.D. Wooley, and D. Zuckerman. Privacy amplification and non-malleable extractors via character sums. In Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science, pages 668-677, 2011.
8. Y. Dodis and D. Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In Proceedings of the 41st Annual ACM Symposium on Theory of Computing, page 601-610, 2009.
9. L. Fortnow and R. Shaltiel. Recent developments in explicit constructions of extractors, 2002. Bulletin of the EATCS 77: pages 67-95, 2002.
10. B. Kanukurthi and L. Reyzin. Key agreement from close secrets over unsecured channels. In EUROCRYPT, pages 206-223, 2009.
11. X. Li. Non-malleable extractors, two-source extractors and privacy amplification. In Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 688-697, 2012.
12. U.M. Maurer and S. Wolf. Privacy amplification secure against active adversaries. In CRYPTO'97, pages 307-321, 1997.

13. U.M. Maurer and S. Wolf. Secret-key agreement over unauthenticated public channels III: Privacy amplification. IEEE Transactions on Information Theory, 49(4): 839-851, 2003.
14. J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. SIAM Journal on Computing, 22(4): 838-856, 1993.
15. N. Nisan and D. Zuckerman. Randomness is linear in space. J. Comput. Syst. Sci., 52(1): 43-52, 1996.
16. A. Rao. An exposition of Bourgain's 2-source extractor. Technical Report TR07-34, ECCC: Electronic Colloquium on Computational Complexity, 2007. http://eccc.hpi-web.de/eccc-reports/2007/TR07-034/index.html.
17. R. Raz. Extractors with weak random seeds. In Proceedings of the 37th Annual ACM Symposium on Theory of Computing, pages 11-20, 2005.
18. R. Renner and S. Wolf. Unconditional authenticity and privacy from an arbitrarily weak secret. In CRYPTO, pages 78-95, 2003.
19. S. Vadhan. Randomness extractors and their many guises: Invited tutorial. In Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science, page 9, 2002.
20. S. Wolf. Strong security against active attacks in information-theoretic secret-key agreement. In Advance in Cryptology ASIACRYPT'98, pages 405-419, 1998.
21. D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In Theory of Computing, pages 103-128, 2007.

A

Theorem 4. *Assuming*

$$0 < 2^{\log 3 - 2\theta + 4m + 8} - 2^{\log 3 - \frac{n}{2} + 4 + \log d - 2\theta + 4m + 8} \leq 2^{2d + 4\theta - 8m - 8 - n + \alpha} - 2^{2d - \frac{n}{2} + 4 + \log d}.$$

Then there exists an explicit $(\alpha, \xi = 2^{\theta})$*-non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$.

Proof. The explicit construction we present is the extractor constructed in [17]. Alon et al. [1] observed that for every k', $N \geq 2$, the sequence of 0-1 random variables $Z_1, \cdots, Z_N$ that is ϵ-biased for linear tests of size k' can be explicitly constructed using $2 \cdot \lceil \log(1/\epsilon) + \log k' + \log\log N \rceil$ random bits. Therefore, let $D = m \cdot 2^d$ and $\epsilon = 2^{-\frac{n}{2}+r}$ with $r = 1 + \log k' + \log\log D$, then we can construct a sequence of 0-1 random variables $Z_1, \cdots, Z_D$ that is ϵ-biased for linear tests of size k' by using n random bits. Let $k' = 8m$. We interpret the set of indices $[D]$ as the set $\{(i,s) : i \in [m], s \in \{0,1\}^d\}$. Define $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ by

$$\mathsf{Ext}(w,s) = Z_{(1,s)}(w) || Z_{(2,s)}(w) \ldots || Z_{(m,s)}(w),$$

where "||" is the concatenation operator.

Let S be a random variable uniformly distributed over $\{0,1\}^d$.

Assume for contradiction that Ext is not a $(\alpha, 2^{\theta})$-non-malleable-extractor. Then there exists a source W of length n with min-entropy α, and an adversarial-function $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ such that

$$\|(\mathsf{Ext}(W,S), \mathsf{Ext}(W, \mathcal{A}(S)), S) - (U_m, \mathsf{Ext}(W, \mathcal{A}(S)), S)\|_1 > \xi = 2^{\theta}.$$

As in [3], we may assume that W is uniformly distributed over a set $W' \subseteq \{0,1\}^n$ of size α.

For every $s \in \{0,1\}^d$, let X_s be the random variable $\mathsf{Ext}(W,s)$. By Lemma 2 and 3, we have

$$\sum_{\emptyset \neq \sigma \subseteq [m], \tau \subseteq [m]} \mathbb{E}_{s \sim S}[bias((X_s)_\sigma \oplus (X_{\mathcal{A}(s)})_\tau)] > 2^\theta.$$

Let $\sigma^*, \tau^* \subseteq [m]$ be the indices of (one of) the largest summands in the above sum. For every $s \in \{0,1\}^d$, let

$$Y_s = (X_s)_{\sigma^*} \oplus (X_{\mathcal{A}(s)})_{\tau^*}.$$

There is a set $S'' \subseteq \{0,1\}^d$ satisfying that $|S''| > \frac{\xi \cdot 2^{d-2}}{2^{mt}(2^m-1)(t+1)^2} = \frac{\xi \cdot 2^{d-2}}{2^{m+2}(2^m-1)}$. The S'' here is the same as that in the proof of Theorem 1 by replacing t there with 1 and the error 2^{-m} there with ξ. Please see [5] for details.

Define a random variable $Y_{S''}$ over $\{0,1\}$ as follows: To sample a bit from $Y_{S''}$, uniformly sample a string s from S'', and then independently sample a string w uniformly from W'. The sampled value is $Y_s(w)$. We have that

$$bias(Y_{S''}) > \frac{\xi}{2^{mt+1}(2^m-1)(t+1)} = \frac{\xi}{2^{m+2}(2^m-1)}.$$

For every $s \in S''$, let

$$Y'_s = \oplus_{i \in \sigma^*} Z_{(i,s)} \oplus (\oplus_{j \in \tau^*} Z_{(j,\mathcal{A}(s))}).$$

Let $t = 1$ in Claim 7.2 of [5], we get the following claim.

Claim 2'. *The set of random variables* $\{Y'_s\}_{s \in S''}$ $\epsilon-$*fools linear tests of size* $\frac{k'}{(t+1)m} = 4$.

We apply Theorem 2 on the random variables $\{Y'_s\}_{s \in S''}$. For simplicity of presentation, we assume $|S''| = 2^{d'}$. By Theorem 2, the distribution of $\mathsf{Ext}^{(1)}(W,S'')$ is $\gamma-$biased for $\gamma = 2^{\frac{n-\alpha}{k}} \cdot [2^{-\frac{d'k}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}$. Let $k = 4$, then $\gamma = 2^{\frac{n-\alpha}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$. We note that $\mathsf{Ext}^{(1)}(W,S'')$ has the same distribution as $Y_{S''}$. In particular, both random variables have the same bias. Therefore, we get

$$2^{\frac{n-\alpha}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} \geq bias(Y_{S''}) > \frac{\xi}{2^{m+2}(2^m-1)}.$$

Moreover, since $2^{d'} = |S''| > \frac{\xi \cdot 2^{d-2}}{2^{m+2}(2^m-1)}$, we have

$$2^{\frac{n-\alpha}{4}} \cdot [\xi^{-2} \cdot 2^{-2d+2m+8} \cdot (2^m-1)^2 \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} > \frac{\xi}{2^{m+2} \cdot (2^m-1)}.$$

Hence,

$$2^{n-\alpha} \cdot [2^{-2\theta} \cdot 2^{-2d+4m+8} \cdot 3 \cdot (1-\epsilon) + \epsilon] > \frac{2^{4\theta}}{2^{8m+8}}.$$

That is,

$$2^{-2d} > \frac{2^{4\theta-8m-8-n+\alpha} - \epsilon}{3(1-\epsilon)2^{-2\theta+4m+8}}$$

with $\epsilon = 2^{-\frac{n}{2}+4+\log d}$, which is in contradiction to the assumption of the theorem. □

Definition 6. (see [5, 8]) In an (n, k, m, δ)-*privacy amplification protocol* (or *information-theoretic key agreement protocol*), Alice and Bob share a weak secret W, and have two candidate keys $r_A, r_B \in \{0,1\}^m \cup \perp$, respectively. For any adversarial strategy employed by Eve, denote two random variables R_A, R_B as the values of the candidate keys r_A, r_B at the conclusion of the protocol execution, and random variable T_E as the transcript of the (entire) protocol execution as seen by Eve. We require that for any weak secret W with min-entropy at least k the protocol satisfies the following three properties:

- **Correctness**: If Eve is passive, then one party reaches the state, the other party reaches the $\mathsf{KeyConfirmed}$ state, and $R_A = R_B$.

- **Privacy**: Denote $\mathsf{KeyDerived}_A$ and $\mathsf{KeyDerived}_B$ as the indicators of the events in which Alice and Bob reach the $\mathsf{KeyDerived}$ state, respectively. Then for any adversarial strategy employed by Eve, if Bob reaches the $\mathsf{KeyDerived}_B$ state during the protocol execution, then $\mathsf{SD}((R_B, T_E), (U_m, T_E)) \leq \delta$; if Alice reaches the $\mathsf{KeyDerived}_A$ state during the protocol execution, then $\mathsf{SD}((R_A, T_E), (U_m, T_E)) \leq \delta$.

- **Authenticity**: Denote $\mathsf{KeyConfirmed}_A$ and $\mathsf{KeyConfirmed}_B$ as the indicators of the events in which Alice and Bob reach the $\mathsf{KeyConfirmed}$ state, respectively. Then, for any adversarial strategy employed by Eve, it holds that

$$Pr[(\mathsf{KeyConfirmed}_A \vee \mathsf{KeyConfirmed}_B) \wedge R_A \neq R_B] \leq \delta.$$