Non-Malleable Extractors with Shorter Seeds and Their Applications

Yanqing Yao[1,2] ** and Zhoujun Li[1,2]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Beijing Key Laboratory of Network Technology, Beihang University, Beijing, China

Abstract. Motivated by the problem of how to communicate over a public channel with an active adversary, Dodis and Wichs (STOC'09) 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, by developing the combination and permutation techniques, we improve the error estimation of the extractor of Raz (STOC'05), which plays an extremely important role in the constraints of the non-malleable extractor parameters including seed length. Then we present improved explicit construction of non-malleable extractors. Though our construction is the same as that given by Cohen, Raz and Segev (CCC'12), the parameters are improved. More precisely, we construct an explicit $(1016, \frac{1}{2})-$non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ with $n = 2^{10}$ and seed length $d = 19$, while Cohen et al. showed that the seed length is no less than $\frac{46}{63} + 66$. Therefore, our method beats the condition "$2.01 \cdot \log n \leq d \leq n$" proposed by Cohen et al., since d is just $1.9 \cdot \log n$ in our construction. We also improve the parameters of the general explicit construction given by Cohen et al. Finally, we give their applications to privacy amplification.

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

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. [12, 14, 22]. In this paper, we focus on the extractors that can be applied to privacy amplification. In this scenario, two parties Alice and Bob share a weakly random secret $W \in \{0,1\}^n$. 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 public seed length and min-entropy of W 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 $R = \mathsf{Ext}(W, S)$ as the nearly uniform secret key [18].

** Most of this work was done while the author visited New York University.

If Eve is active (i.e., it may change the messages in arbitrary ways), some protocols have been proposed to achieve this goal [4, 6–9, 13–15, 21, 23].

As a major progress, Dodis and Wichs [9] 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 [1] W with min-entropy α and uniformly random seed S, and outputs a string which is $\gamma-$close to uniform (see Definition 1), 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 $(\alpha, 2\gamma)-$non-malleable extractors exist as long as $\alpha > 2m + 3\log\frac{1}{\gamma} + \log d + 9$ and $d > \log(n - \alpha + 1) + 2\log\frac{1}{\gamma} + 7$. The first explicit non-malleable extractor was constructed by Dodis, Li, Wooley and Zuckerman [8]. It works for any weakly random input source with the min-entropy $\alpha > \frac{n}{2}$ 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.

Li [14] proposed that $(\alpha, 2\gamma)$-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}$, where $\alpha = (\frac{1}{2} - \delta) \cdot n$ and $d = O(\log n + \log(1/\gamma))$ for any constant $\delta > 0$, can be constructed as follows: the seed S is encoded using the parity check matrix of a BCH code, and then the output is the inner product function of the encoded source and the encoded seed over $\mathbb{F}_2$. Dodis and Yu [11] observed that for 4-wise independent hash function family $\{h_w : \{0,1\}^d \to \{0,1\}^m \mid w \in \{0,1\}^n\}$, $\mathsf{nmExt}(w, s) = h_w(s)$ is a $(\alpha, 2\sqrt{2^{n-\alpha-d}})$-non-malleable extractor. In 2012, an alternative explicit construction based on the extractor of Raz [20] was given by Cohen et al. [6]. Without using any conjecture, their construction works for any weakly random source with the min-entropy $\alpha = (\frac{1}{2} + \delta) \cdot n$ and uniformly random seed of length $d \geq \frac{23}{\delta} \cdot m + 2\log n$ (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 [20], while the error [2] estimation in that construction is too rough. Furthermore, though one main purpose of [6] is to shorten the length of the seed, the lower bound on the seed length is still not optimal.

OUR CONTRIBUTIONS AND TECHNIQUES.

- By developing the combination and permutation techniques, we improve the error estimation of Raz's extractor in STOC'05 [20], a special case of which was used by Cohen et al. in CCC'12 [6]. For simplicity, denote γ_1 as the error of the extractor in [6], 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 [6] under the assumption that $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k$ and $0 < \delta \leq \frac{1}{2}$ (see Lemma 1). 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 [6, 20] which bounds the error didn't capture the essence of the biased sequence for linear tests (see Definition 2). In this paper, we propose another partition method and give a better bound on the sum by employing the combination and permutation formulas. In particular, the combination and permutation techniques (see Proposition 1) may be useful in future works. Correspondingly, the error is $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot [2^{-\frac{dk}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}$ (see Theorem 2). 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

[1] When we say a source in this paper, we mean a random variable.

[2] The concept of the error of seeded extractor can be seen in Definition 1.

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 $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)n}{4}} \cdot [2^{-2d} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$.

- Note that the error estimation of the Raz's extractor impacts greatly on the constraints of the parameters including the seed length, the weak source's min-entropy and the error [3] of the non-malleable extractor. Based on the above 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 [6] in the sense that the seed length is shorter. More concretely, we present an explicit $(1016, \frac{1}{2})-$non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ with $n = 1024$ and $d = 19$, which beats the condition "$2.01 \cdot \log n \leq d \leq n$" in [6], since seed length d is just $1.9 \cdot \log n$ in our construction while it is no less than $\frac{46}{63} + 66$ according to [6]. Moreover, we improve the parameters of the general explicit construction given by Cohen et al.
- We show how our non-malleable extractors are applied to privacy amplification.

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 [6]. In Section 5, we show how the non-malleable extractors are applied to privacy amplification. Section 6 concludes the paper.

2 Preliminaries

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 [6]. 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, $\mathsf{SD}(f(X), f(Y)) \leq \mathsf{SD}(X,Y)$. Denote $\mathsf{SD}((X_1,X_2),(Y_1,Y_2) \mid Z)$ as the abbreviation of $\mathsf{SD}((X_1,X_2,Z),(Y_1,Y_2,Z))$.

The *min-entropy* of variable W is $H_\infty(W) = -\log \max_w Pr(W = w)$. W over $\{0,1\}^n$ is called an (n,α)*-source* if $H_\infty(W) \geq \alpha$. 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 [5] observed that the distribution of any (n,α)-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. *We say that the distribution X is ϵ-close to the distribution Y if $\|X - Y\|_1 = \sum_s |\Pr[X = s] - \Pr[Y = s]| \leq \epsilon$ [4]. A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is an $(\alpha,\gamma)-$seeded extractor if for every (n,α)-source W and an independent uniformly random variable S (called seed) over $\{0,1\}^d$, the distribution of $\mathsf{Ext}(W,S)$ is γ-close to U_m. γ is called the error of the seeded extractor. A seeded extractor is a strong $(\alpha,\gamma)-$extractor if for W and S as above, $(\mathsf{Ext}(W,S),S)$ is γ-close to (U_m,U_d).*

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

[3] The concept of the error of non-malleable extractor can be seen in Definition 3.

[4] In other papers (e.g., [9, 11, 14, 24]), X is ϵ-close to Y if $\frac{1}{2}\|X - Y\|_1 = \frac{1}{2}\sum_s |\Pr[X = s] - \Pr[Y = s]| \leq \epsilon$. To keep consistency, Definition 1 holds throughout this paper.

$|\tau| \le k$, 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 [20] 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 [20].

Cohen et al. [6] 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 [6]) We say that a function $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ is an adversarial function, if for every $s \in \{0,1\}^d$, $f(s) \ne s$ holds. A function $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (α, γ)-non-malleable extractor if for every (n, α)-source W, independent uniformly random variable S, and every adversarial function $\mathcal{A}$,*

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

γ is called the error of the non-malleable extractor.

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

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

Recall that the main theorem about the explicit construction of non-malleable extractors proposed in [6] is as follows.

Theorem 1. (see [6]) *For any integers n, d, and m, and for any $0 < \delta \le \frac{1}{2}$ such that $d \ge \frac{23}{\delta} \cdot m + 2 \log n$, $n \ge \frac{160}{\delta} \cdot m$, and $\delta \ge 10 \cdot \frac{\log(nd)}{n}$, there exists an explicit $((\frac{1}{2} + \delta) \cdot n, 2^{-m})$-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$.*

3 Error Estimation of Raz's Extractor and its Improvement

In this section, we first recall the central lemma used in [6], which is a special case about the error estimation of Raz's Extractor [20]. 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 [6] is below, the proof of which is essentially the same as that in [20]. It can be considered as a special case of Raz's Extractor [20].

Lemma 1. *Let $D = 2^d$. Let $Z_1, \ldots, Z_D$ 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 $Z_1, \ldots, Z_D$. Then, for any $0 < \delta \le \frac{1}{2}$ and even integer $k \le k'$ s.t. $k \cdot (\frac{1}{\epsilon})^{\frac{1}{k}} \le D^{\frac{1}{2}}$, $\mathsf{Ext}^{(1)}$ is a $((\frac{1}{2}+\delta)\cdot n, \gamma_1)$-seeded-extractor, with $\gamma_1 = (\epsilon \cdot 2^{(\frac{1}{2}-\delta)n+1})^{\frac{1}{k}}$.*

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 γ_1-close to uniform. As in [5], 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 $s_1, \ldots, s_r \in \{0,1\}^d$,*

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

Proof.

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

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

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

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

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

The sum over $s_1,\dots,s_k \in \{0,1\}^d$ is broken into two sums. The first sum is over $s_1,\dots,s_k \in \{0,1\}^d$ such that in each summand, at least one s_j is different than all other elements in the sequence $s_1,\dots,s_k$ [5], and the second sum is over $s_1,\dots,s_k \in \{0,1\}^d$ such that in each summand every s_j is identical to at least one other element in the sequence $s_1,\dots,s_k$. 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,

$$2^{(\frac{1}{2}+\delta)n}\cdot 2^{d\cdot k}\cdot\gamma(W,S)^k \le 2^n\cdot\epsilon\cdot 2^{d\cdot k}+2^n\cdot 2^{d\cdot\frac{k}{2}}\cdot(\frac{k}{2})^k \le 2\cdot 2^n\cdot\epsilon\cdot 2^{d\cdot k},$$

where the last inequality follows by the assumption that $k\cdot(1/\epsilon)^{1/k}\le D^{\frac{1}{2}}$. That is, $\gamma(W,S)\le(\epsilon\cdot 2^{(\frac{1}{2}-\delta)n+1})^{\frac{1}{k}}$. □

The above partition method about the sum over $s_1,\dots,s_k\in\{0,1\}^d$ is not optimal, since it doesn't capture the essence of random variable sequence that is biased for linear tests (i.e., $Z_1,\dots,Z_{2^d}$ is called ϵ-biased for linear tests of size k if for any nonempty $\tau\subseteq[2^d]$ with $|\tau|\le k$, the random variable $Z_\tau=\oplus_{i\in\tau}Z_i$ is $\epsilon-$biased). Moreover, the bounds on the number of summands in the two sums are too large. The same problem exists in [20].

In fact, when every s_j is identical to at least one other element in the sequence $s_1,\dots,s_k$ under the assumption that at least one s_j appears odd times in the sequence $s_1,\dots,s_k$, 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, …, Z_D are 0-1 random variables that are ϵ-biased for linear tests of size k'. However, in this case the upper bound on 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 [6, 20].

3.2 Improvement for the Error Estimation of Raz's Extractor

We improve the error estimation of Raz's extractor as follows. Unlike [6, 20], we present another partition method of the sum in the following proof. The combination and permutation formulas are exploited to show a tight bound on the sum. Correspondingly, the error can be reduced.

Proposition 1. *Consider fixed positive numbers k and d. Assume that a sequence $s_1,\dots,s_k$ satisfies the following two conditions: (1) for every $i\in[k]$, $s_i\in\{0,1\}^d$, and (2) for every $j\in[k]$, s_j appears even times in the sequence $s_1,\dots,s_k$. Then the number of such sequences $s_1,\dots,s_k$ is $2^{\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1$.*

Proof. Denote C_r^l as the number of possible combinations of r objects from a set of l objects. Then $C_r^l=\frac{l!}{r!(l-r)!}=\frac{l(l-1)(l-2)\cdots(l-r+1)}{r!}$. Denote P_r^l as the number of possible permutations of r objects from a set of l objects. Then $P_r^l=\frac{l!}{(l-r)!}=l(l-1)(l-2)\cdots(l-r+1)$. Hence the number of the corresponding sequences is

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

[5] In this paper, two elements s_i and s_j in the sequence $s_1,\dots,s_k$, where $i\neq j$, might represent the same string.

□

Theorem 2. *Let $D = 2^d$. Let $Z_1, \dots, Z_D$ be 0-1 random variables that are ϵ-biased for linear tests of size k' that are constructed using n random bits. Define $\mathit{Ext}^{(1)}$: $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ by $\mathit{Ext}^{(1)}(w,s) = Z_s(w)$, that is, $\mathit{Ext}^{(1)}(w,s)$ is the random variable Z_s, when using w as the value of the n bits needed to produce $Z_1, \dots, Z_D$. Then, for any $0 < \delta \le \frac{1}{2}$ and any even integer $k \le k'$, $\mathit{Ext}^{(1)}$ is a $((\frac{1}{2}+\delta)\cdot n, \gamma_2)$-seeded-extractor, where $\gamma_2 = 2^{\frac{(\frac{1}{2}-\delta)\cdot n}{k}} \cdot [2^{-\frac{dk}{2}} \cdot (k-1)\cdot(k-3)\cdot\dots\cdot 1\cdot(1-\epsilon)+\epsilon]^{\frac{1}{k}}$.*

Proof. We improve the proof by proposing another method for partitioning the sum $\sum\limits_{s_1,\dots,s_k\in\{0,1\}^d}\sum\limits_{w\in\{0,1\}^n}\prod\limits_{j=1}^{k} e(w,s_j)$ into two sums. The first sum is over $s_1,\dots,s_k \in \{0,1\}^d$ such that in each summand, at least one s_j appears odd times in the sequence $s_1,\dots,s_k$, and the second sum is over $s_1,\dots,s_k \in \{0,1\}^d$ such that in each summand every s_j appears even times in the sequence $s_1,\dots,s_k$. By Proposition 1, the number of summands in the second sum is $2^{\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1$, and each summand is 2^n. Therefore, the number of summands in the first sum is $2^{dk} - 2^{\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1$, and by Claim 1 each summand is bounded by $2^n\cdot\epsilon$. Hence, $2^{(\frac{1}{2}+\delta)\cdot n}\cdot 2^{d\cdot k}\cdot\gamma(W,S)^k \le 2^n\cdot[2^{\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1]+2^n\cdot\epsilon\cdot[2^{dk}-2^{\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1]$. Correspondingly,

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

That is, $\gamma(W,S) \le 2^{\frac{(\frac{1}{2}-\delta)\cdot n}{k}}\cdot[2^{-\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1\cdot(1-\epsilon)+\epsilon]^{\frac{1}{k}}$. □

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 2. *$(k-1)\cdot(k-3)\cdot\dots\cdot 1 \le (\frac{k}{2})^k$ for any positive even integer k, and "=" holds iff $k=2$. Furthermore, $\lim\limits_{k\to\infty}\frac{(k-1)\cdot(k-3)\cdot\dots\cdot 1}{2^{\frac{1}{2}}\cdot(\frac{k}{e})^{\frac{k}{2}}} = 1$.*

Proof. When $k=2$, it's trivial that $(k-1)\cdot(k-3)\cdot\dots\cdot 1 = (\frac{k}{2})^k$. 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,

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

From the Stirling's Formula, we have $\lim\limits_{k\to\infty}\frac{k!}{\sqrt{2\pi k}(\frac{k}{e})^k} = 1$. Therefore,

$$\lim_{k\to\infty}\frac{(k-1)\cdot(k-3)\cdot\dots\cdot 1}{2^{\frac{1}{2}}\cdot(\frac{k}{e})^{\frac{k}{2}}} = \lim_{k\to\infty}[\frac{k!}{\sqrt{2\pi k}\cdot(\frac{k}{e})^k}\cdot\frac{\sqrt{2\pi\cdot\frac{k}{2}}\cdot(\frac{k}{2e})^{\frac{k}{2}}}{(\frac{k}{2})!}] = 1.$$

□

The error estimation of the extractor in Theorem 1 is better than that in Lemma 1. Recall that in Theorem 1, we have

$$\begin{aligned}\gamma_2 &= 2^{\frac{(\frac{1}{2}-\delta)n}{k}}\cdot[2^{-\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1\cdot(1-\epsilon)+\epsilon]^{\frac{1}{k}}\\ &= 2^{\frac{(\frac{1}{2}-\delta)n}{k}}\cdot\{2^{-\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1+[1-2^{-\frac{dk}{2}}\cdot(k-1)\cdot(k-3)\cdot\dots\cdot 1]\cdot\epsilon\}^{\frac{1}{k}},\end{aligned}$$

while in Lemma 1, we have $\gamma_1 = 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot (2\epsilon)^{\frac{1}{k}}$ in [6] under the assumption that $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k$ and $0 < \delta \leq \frac{1}{2}$.

In general, 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 $\gamma_1 > \gamma_2$. In particular, when k is large enough, from Proposition 2, we get that $(k-1) \cdot (k-3) \cdot \cdots \cdot 1 \approx 2^{\frac{1}{2}} \cdot (\frac{k}{e})^{\frac{k}{2}}$. Therefore,

$$\gamma_2 \approx 2^{\frac{(\frac{1}{2}-\delta)n}{k}} \cdot \{2^{-\frac{dk}{2}} \cdot 2^{\frac{1}{2}} \cdot (\frac{k}{e})^{\frac{k}{2}} + [1 - 2^{-\frac{dk}{2}} \cdot 2^{\frac{1}{2}} \cdot (\frac{k}{e})^{\frac{k}{2}}] \cdot \epsilon\}^{\frac{1}{k}}.$$

Correspondingly, $\epsilon \geq 2^{-\frac{dk}{2}} \cdot k^k > 2^{-\frac{dk}{2}} \cdot 2^{\frac{1}{2}} \cdot (\frac{k}{e})^{\frac{k}{2}}$. Hence, $\gamma_1 > \gamma_2$.

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

Remark 2. Noted that 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.

Remark 3. 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.

3.4 Important Role in Improving the Seed Length of Non-Malleable Extractors

It should be noticed that the error of the non-malleable extractor in Theorem 1 given by Cohen et al. [6] relies on some constrained parameters. The main idea of the proof about Theorem 1 given by Cohen et al. [6] is as follows. Assume 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 2, 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 $\gamma_2 < A$. See the proofs of Theorems 3 and 4 below for details. Correspondingly, the seed length may be further shortened.

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

In this section, we improve the parameters of the explicit construction of non-malleable extractors by Cohen et al. in [6]. The seed length here is shorter than that in Theorem 1.

We first review two lemmas that will be used later.

Lemma 2. (see [6]) *Let X be a random variable over $\{0,1\}^m$. Let Y, S be two random variables. Then,*

$$\|(X,Y,S) - (U_m,Y,S)\|_1 = \mathbb{E}_{s \sim S}[\|(X,Y,S)|_{S=s} - (U_m,Y,S)|_{S=s}\|_1].$$

Lemma 3. (see [6]) *Let X, Y be random variables over $\{0,1\}^m$ and $\{0,1\}^n$ respectively. 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 i-th bit of X, Y_j is the jth bit of Y, $X_\sigma = \oplus_{i \in \sigma} X_i$, and $Y_\tau = \oplus_{j \in \tau} Y_j$.*

In what follows, we show a specific explicit construction of a non-malleable extractor such that it is an improvement of [6] in the sense that the seed length is shorter.

Theorem 3. *There exists an explicit $(1016, \frac{1}{2})$-non-malleable extractor $\mathsf{Ext} : \{0,1\}^{1024} \times \{0,1\}^{19} \to \{0,1\}$.*

Proof Idea. *We borrow the reductio ad absurdum approach in the proof of Theorem 1. The proof sketch is as follows. Assume by contradiction that Ext is not non-malleable. Then*

Phase 1: There must exist a weak source W with min-entropy at least α and an adversarial function $\mathcal{A}$ such that the statistical distance between $(\mathsf{Ext}(W,S),$ $\mathsf{Ext}(W, \mathcal{A}(S)), S)$ and $(U_1, \mathsf{Ext}(W, \mathcal{A}(S)), S)$ has a certain lower bound. Then there exists $S \subseteq \{0,1\}^d$ s.t. for every $s \in S$, $Y_s = \mathsf{Ext}(W,s) \oplus \mathsf{Ext}(W, \mathcal{A}(s))$ is biased. Consider the directed graph $G = (S \cup \mathcal{A}(s), E)$ with $E = \{(s, \mathcal{A}(s) : s \in S\}$, where G might contains cycles. By employing a lemma about graph as shown in [6], we can find a subset $S' \subseteq S$ s.t. the induced graph of G by $S' \cup \mathcal{A}(S')$ is acyclic.

Phase 2: We prove that the set of variables $\{Y_s\}_{s \in S'}$ is ϵ-biased for linear tests of size at most $k/2$. Consider the extractor of Raz built on the variables $\{Y_s\}_{s \in S'}$. It's a good seeded-extractor, which yields a contradiction.

Phase 1 of the proof is almost the same as that in [6]. Phase 2 jumps out of the idea in [6]. 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 $\frac{\lceil 128\delta \rceil}{2}$, where δ can be seen in Theorem 1. Therefore the extractor error can be simplified and we don't need to prove $k \cdot (\frac{1}{\epsilon})^{\frac{1}{k}} \leq (2^d)^{\frac{1}{2}}$ as shown in Lemma 1.

Proof. The explicit construction we present is the extractor constructed in [20]. Alon et al. [1] observed that for every k', $N \geq 2$, the sequence of 0-1 random variables $Z_1, \ldots, Z_N$ 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 $D = 2^{19}$ and $\epsilon = 2^{-\frac{1024}{2} + r}$ with $r = 1 + \log k' + \log 19$, then we can construct a sequence of 0-1 random variables $Z_1, \ldots, Z_{2^{19}}$ that is ϵ-biased for linear tests of size k' using n random bits. Let $k' = 8$. Define $\mathsf{Ext} : \{0,1\}^{1024} \times \{0,1\}^{19} \to \{0,1\}$ by $\mathsf{Ext}(w,s) = Z_s(w)$.

Let S be a random variable uniformly distributed over $\{0,1\}^{19}$.

Assume for contradiction that Ext is not a $(1016, \frac{1}{2})$-non-malleable-extractor. Then there exists a source W of length 1024 with min-entropy 1016, and an adversarial function $\mathcal{A} : \{0,1\}^{19} \to \{0,1\}^{19}$ such that

$$\|(\mathsf{Ext}(W,S), \mathsf{Ext}(W, \mathcal{A}(S)), S) - (U_1, \mathsf{Ext}(W, \mathcal{A}(S)), S)\|_1 > \frac{1}{2}.$$

As in [5], suppose W is uniformly distributed over a set $W' \subseteq \{0,1\}^{1024}$ of size 2^{1016}.

For every $s \in \{0,1\}^{19}$, let X_s be the random variable $\mathsf{Ext}(W,s)$. By Lemmas 2 and 3, we have

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

Let $\sigma^*, \tau^* \subseteq [1]$ be the indices of (one of) the largest summands in the above sum. For every $s \in \{0,1\}^{19}$, let $Y_s = (X_s)_{\sigma^*} \oplus (X_{\mathcal{A}(s)})_{\tau^*}$.

There is a set $S'' \subseteq \{0,1\}^{19}$ satisfying that

$$|S''| > \frac{\xi \cdot 2^{19-2}}{2(1+1)^2} = 2^{13}.$$

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 $\frac{1}{2}$. Please see [6] 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 $bias(Y_{S''}) > \frac{\frac{1}{2}}{2^{1+1}(2-1)(1+1)} = \frac{1}{2^4}$. 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 [6], we get the following claim.

Claim 2. *The set of random variables $\{Y'_s\}_{s \in S''}$ $\epsilon-$fools linear tests of size 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 γ_2-biased for $\gamma_2 = 2^{\frac{8}{k}} \cdot [2^{-\frac{d'k}{2}} \cdot (k-1) \cdot (k-3) \cdot \dots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}$. Let $k = \frac{k'}{2} = 4$, then $\gamma_2 = 2^{\frac{8}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$. 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

$$2^{\frac{8}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} \geq bias(Y_{S''}) > \frac{1}{2^4},$$

Moreover, since $2^{d'} = |S''| > 2^{13}$, we have

$$2^2 \cdot [4 \cdot 2^{-28} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} > 2^2 \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}} > \frac{1}{2^4}.$$

That is,

$$2^{-38} > \tfrac{2^{-4} \cdot 2^{-20} - \epsilon}{3(1-\epsilon) \cdot 2^{12}}, \qquad (a)$$

where $\epsilon = 2^{-516+r}$ and $r = 4 + \log 19$.

On the other hand, we have $2^{-38} < \frac{2^{-4} \cdot 2^{-20} - \epsilon}{3(1-\epsilon) \cdot 2^{10} \cdot 2^2}$, 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 $0 < \delta \leq \frac{1}{2}$. However, in the above construction, we have $d = 1.9 \log n$. We compare them in detail as follows.

Let $n = 2^{10}$, $m = 1$, and $\delta = \frac{63}{128}$ in Theorem 1, then it can be easily verified that $n \geq \frac{160}{\delta} \cdot m$. To construct an explicit $((\frac{1}{2} + \delta) \cdot n, 2^{-m})$-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m$ (that is, an explicit $(1016, \frac{1}{2})$-non-malleable extractor nmExt), by Theorem 1, the seed length d should satisfy $d \geq \frac{23}{\delta} \cdot m + 2 \log n = \frac{46}{63} + 66$. 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 [6].

Using the extractor with improved error estimation (see Theorem 2), we can also improve the parameters of the explicit non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m$ constructed by Cohen et al. [6] below.

Theorem 4. *Assume that*

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

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

The proof is similar to that of Theorem 3. Please see Appendix A for details.

Due to the analysis of Section 3.4, we conclude that the above theorem is really an improvement in the sense that the seed length here is shorter. Though the constrains on the parameters in Theorem 4 are complex, we show some simplification in Appendix B. How to further simplify the constraints is an open problem.

5 Application to Privacy Amplification

In this section, we show how the non-malleable extractor is applied to the privacy amplification protocol [8, 9] (also known as an information-theoretic key agreement protocol), the formal concept of which can be seen in Appendix C.

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.

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 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 KeyConfirmed state, and output $R_A = \mathsf{Ext}(W, S_0')$.		

Table 1. The Dodis-Wichs privacy amplification protocol.

Assume that 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 [6, 8]. 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'. 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.

Theorem 6. (see [6, 9]) *Assume* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^{d_1} \rightarrow \{0,1\}^{m_1}$ *is a* $(\alpha, \gamma_{\mathsf{nmExt}})$*-non-malleable extractor,* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^{d_2} \rightarrow \{0,1\}^{m_2}$ *is a strong* $(\alpha - (d_1 + m_1) - \log \frac{1}{\epsilon'}, \gamma_{\mathsf{Ext}})$*-extractor, and* $\{\mathsf{MAC}_r : \{0,1\}^{d_2} \rightarrow \{0,1\}^{\tau}\}_{r \in \{0,1\}^{m_1}}$ *is a* $\varepsilon_{\mathsf{MAC}}$*-secure message authentication code. Then for any integers* n *and* $\alpha \leq n$*, the protocol in Table 1 is a 2-round* (n, α, m, η)*-privacy amplification protocol, with communication complexity* $d_1 + d_2 + \tau$ *and* $\eta = \max\{\epsilon' + \gamma_{\mathsf{Ext}}, \gamma_{\mathsf{nmExt}} + \varepsilon_{\mathsf{MAC}}\}$.

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 using the combination and permutation techniques. Based on the improvement, we presented an improved explicit construction of non-malleable extractors with shorter seed length. Similar to [6], our construction is also based on biased variable sequence for linear tests. However, our parameters are improved. More precisely, we presented an explicit $(1016, \frac{1}{2})-$non-malleable extractor $\mathsf{nmExt} : \{0,1\}^{1024} \times \{0,1\}^d \rightarrow \{0,1\}$ with seed length 19, while it is no less than $\frac{46}{63} + 66$ according to Cohen et al. in CCC'12 [6]. We also improved the parameters of the general explicit construction of non-malleable extractors proposed by Cohen et al. and analyzed the simplification of the constraints on the parameters (see Appendix B for details). How to further simplify the constraints is an open problem. Finally, we showed their applications to 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 in part by High Technology Research and Development Program of China under grant No. 2015AA016004, NSFC (Nos. 61170189, 61370126, 61202239), the Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2015ZX-16), the Fund of the China Scholarship Council under grant No. 201206020063, and the Fundamental Research Funds for the Central Universities.

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. M. Cheraghchi, V. Guruswami. Non-malleable Coding against Bit-Wise and Split-State Tampering. TCC 2014, pages 440-464.
4. N. Chandran, B. Kanukurthi, R. Ostrovsky, and L. Reyzin. Privacy amplification with asymptotically optimal entropy loss. STOC 2010, pages 785-794.
5. 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.
6. G. Cohen, R. Raz, and G. Segev. Non-malleable Extractors with Short Seeds and Applications to Privacy Amplification. CCC 2012, pages 298-308.

7. Y. Dodis, J. Katz, L. Reyzin, and A. Smith. Robust fuzzy extractors and authenticated key agreement from close secrets. CRYPTO 2006, pages 232-250.
8. Y. Dodis, X. Li, T.D. Wooley, and D. Zuckerman. Privacy amplification and non-malleable extractors via character sums. FOCS 2011, pages 668-677.
9. Y. Dodis and D. Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. STOC 2009, page 601-610.
10. S. Dziembowski, K. Pietrzak, and D. Wichs. Non-malleable codes. In Proceedings of Innovations in Computer Science (ICS 2010), pages 434-452.
11. Y. Dodis, Y. Yu. Overcoming Weak Expectations. TCC 2013, pages 1-22.
12. L. Fortnow and R. Shaltiel. Recent developments in explicit constructions of extractors, 2002. Bulletin of the EATCS 77: pages 67-95, 2002.
13. B. Kanukurthi and L. Reyzin. Key agreement from close secrets over unsecured channels. EUROCRYPT 2009, pages 206-223.
14. X. Li. Non-malleable extractors, two-source extractors and privacy amplification. FOCS 2012, pages 688-697.
15. U.M. Maurer and S. Wolf. Privacy amplification secure against active adversaries. CRYPTO 1997, pages 307-321.
16. 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.
17. J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. SIAM Journal on Computing, 22(4): 838-856, 1993.
18. N. Nisan and D. Zuckerman. Randomness is linear in space. J. Comput. Syst. Sci., 52(1): 43-52, 1996.
19. A. Rao. An exposition of Bourgain's 2-source extractor. Technical Report TR07-34, ECCC, 2007. http://eccc.hpi-web.de/eccc-reports/2007/TR07-034/index.html.
20. R. Raz. Extractors with weak random seeds. STOC 2005, pages 11-20.
21. R. Renner and S. Wolf. Unconditional authenticity and privacy from an arbitrarily weak secret. CRYPTO 2003, pages 78-95.
22. S. Vadhan. Randomness extractors and their many guises: Invited tutorial. FOCS 2002, page 9.
23. S. Wolf. Strong security against active attacks in information-theoretic secret-key agreement. ASIACRYPT 1998, pages 405-419.
24. D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In Theory of Computing 2007, pages 103-128.

A Proof of Theorem 4

Proof. The explicit construction we present is the extractor constructed in [20]. Alon et al. [1] observed that for every k', $N \geq 2$, the sequence of 0-1 random variables $Z_1, \ldots, Z_N$ 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 $Z_1, \ldots, Z_D$ that is ϵ-biased for linear tests of size k' 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 $\mathsf{Ext}(w,s) = Z_{(1,s)}(w) \cdots || Z_{(m,s)}(w)$, 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

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

As in [5], suppose W is uniformly distributed over $W' \subseteq \{0,1\}^n$ of size 2^{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

$$|S''| > \frac{2^\theta \cdot 2^{d-2}}{2^{mt}(2^m-1)(t+1)^2} = \frac{2^\theta \cdot 2^{d-2}}{2^{m+2}(2^m-1)}.$$

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 2^θ. Please see [6] 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

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

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 [6], we get the following claim.

Claim 2'. *The set $\{Y'_s\}_{s \in S''}$ $\epsilon-$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 γ_2-biased for $\gamma_2 = 2^{\frac{n-\alpha}{k}} \cdot [2^{-\frac{d'k}{2}} \cdot (k-1) \cdot (k-3) \cdot \cdots \cdot 1 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{k}}$. Let $k = 4$, then $\gamma_2 = 2^{\frac{n-\alpha}{4}} \cdot [2^{-2d'} \cdot 3 \cdot (1-\epsilon) + \epsilon]^{\frac{1}{4}}$. 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

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

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

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

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. □

B Analysis of the assumption in Theorem 4

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

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

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* $\Leftrightarrow 0 < 2^{A'} - 2^{B'} \leq 2^{C'} - 2^{D'}$. We discuss what happens under the assumption (b) in three cases as follows.

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

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

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

Let $\rho' \geq -16$. 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 $\epsilon \in (0, 1 - 2^{-16}]$, $A' \geq C'$, and $B' \geq D'$, (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. Assume that $A' \geq C'$ and $B' < D'$, then it's in contradiction to the inequality (b).

Case 3. Assume that $A' < C'$, 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 corollary.

Corollary. *Assume that* $\epsilon = 2^{-\frac{n}{2}+4+\log d} \in (0, 1 - 2^{-16}]$ *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 $(\alpha, 2^{\theta})$*-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

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

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 $4\theta - 8m - 12 - \frac{n}{2} + \alpha > \log d$.

C The concept of privacy amplification protocol

Definition 7. (see [6, 9]) In an (n, α, m, η)*-privacy amplification protocol* (or *information-theoretic key agreement protocol*), Alice and Bob share a weak secret W, and have two candidate keys $r_A, r_B \in \{0,1\}^m \cup \bot$, 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 α 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 during the protocol execution, for any adversarial strategy employed by Eve, if Bob reaches the $\mathsf{KeyDerived}_B$ state then $\mathsf{SD}((R_B, T_E), (U_m, T_E)) \leq \eta$; if Alice reaches the $\mathsf{KeyDerived}_A$ state, then $\mathsf{SD}((R_A, T_E), (U_m, T_E)) \leq \eta$.

- **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

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