Generalized LRS Estimator for Min-entropy Estimation

Jiheon Woo, Chanhee Yoo, Young-Sik Kim, Yuval Cassuto, and Yongjune Kim

Abstract—The min-entropy is a widely used metric to quantify the randomness of generated random numbers, which measures the difficulty of guessing the most likely output. It is difficult to accurately estimate the min-entropy of a non-independent and identically distributed (non-IID) source. Hence, NIST Special Publication (SP) 800-90B adopts ten different min-entropy estimators and then conservatively selects the minimum value among ten min-entropy estimates. Among these estimators, the longest repeated substring (LRS) estimator estimates the collision entropy instead of the min-entropy by counting the number of repeated substrings. Since the collision entropy is an upper bound on the min-entropy, the LRS estimator inherently provides overestimated outputs. In this paper, we propose two techniques to estimate the min-entropy of a non-IID source accurately. The first technique resolves the overestimation problem by translating the collision entropy into the min-entropy. Next, we generalize the LRS estimator by adopting the general Rényi entropy instead of the collision entropy (i.e., Rényi entropy of order two). We show that adopting a higher order can reduce the variance of min-entropy estimates. By integrating these techniques, we propose a generalized LRS estimator that effectively resolves the overestimation problem and provides stable min-entropy estimates. Theoretical analysis and empirical results support that the proposed generalized LRS estimator improves the estimation accuracy significantly, which makes it an appealing alternative to the LRS estimator.

I. INTRODUCTION

Random numbers are essential for generating cryptographic information such as secret keys, nonces, salt values, etc. The security of cryptographic systems crucially relies on the randomness of the generated random numbers [1]–[4]. Hence, it is critical to quantify the randomness of the generated numbers accurately. Among several ways to quantify randomness of generated random numbers, entropies are widely used metrics in standards such as AIS.31 [5], NIST Special Publication (SP) 800-22 [6], and NIST SP 800-90B [1].

There are several kinds of entropies such as Shannon entropy, Rényi entropy, and min-entropy. Among them, the min-entropy is a well-justified metric in cryptographic applications [1], [3] since the min-entropy measures the difficulty of guessing the most likely output. Furthermore, the min-entropy is a lower bound on the Shannon entropy and the Rényi entropy, i.e., one of the most conservative metrics.

For independent and identically distributed (IID) sources, the min-entropy can be readily estimated by the empirical estimator [1]. However, it is difficult to estimate the min-entropy of non-IID sources accurately. Hence, NIST SP 800-90B proposes ten different min-entropy estimators for non-IID sources (see Table I). These estimators independently perform their own estimations based on different statistics of the examined non-IID sources. Then, NIST SP 800-90B conservatively selects the minimum among these ten different values as the final estimate of min-entropy [1], [7].

Among these ten min-entropy estimators, the longest repeated substring (LRS) estimator estimates the collision entropy (the Rényi entropy of order two) based on the number of repeated substrings, i.e., collision counts [1]. Since the collision entropy is an upper bound on the min-entropy, the LRS estimator overestimates the min-entropy, which violates the conservative estimation of NIST SP 800-90B. NIST SP 800-90B selects the minimum among ten estimates as the final min-entropy estimate; hence, the overestimated value by the LRS estimator would typically not affect the final estimate, which could undermine the justification to include the LRS estimator in NIST SP 800-90B [8].

In this paper, we propose two techniques to amend the LRS estimator for accurate min-entropy estimation. The first technique resolves the overestimation problem by enabling the estimation of the min-entropy instead of the collision entropy. The proposed technique leverages the inequality of [9, Theorem 6], which characterizes the relation between the min-entropy and the Rényi entropy. For this technique, we show that the proposed estimator is almost unbiased for binary sources, which are the most common sources. Next, we generalize the proposed estimator by parameterizing the order, i.e., \( \alpha \) of the Rényi entropy. The Rényi entropy is a quantity that generalizes various notions of entropy by parameterizing...
the order $\alpha$ [10]. By adopting a higher order $\alpha$ than two of the collision entropy, the variance of min-entropy estimates can be reduced, which leads to more stable estimates. We analytically show that the variance of estimates decreases with $\alpha$, although the reduction of the variance diminishes as $\alpha$ increases.

By integrating these two techniques, we propose a generalized LRS estimator that improves the estimation accuracy by twofold: 1) the bias is reduced by resolving the overestimation problem of the LRS estimator; 2) the variance of the min-entropy estimates is reduced by adopting the higher order of the Rényi entropy. Theoretical analysis and empirical results support that the generalized LRS estimator significantly improves the estimation accuracy of the LRS estimator, although both the LRS estimator and the proposed estimator rely on the same statistics, i.e., counts of repeated substrings. We believe that the generalized LRS estimator is an appealing alternative to the LRS estimator of NIST SP 800-90B since the generalized LRS estimator outputs a more accurate and stable min-entropy estimate from the same statistics of a given sequence.

The rest of this paper is organized as follows. Section II briefly explains the several types of entropies and the LRS estimator of NIST SP 800-90B. Section III presents the improved LRS estimator that accurately estimates the min-entropy. Section IV proposes the generalized LRS estimator that enables more stable estimation. Section V provides numerical results and Section VI concludes.

II. PRELIMINARIES: ENTROPIES AND LRS ESTIMATOR

A. Entropies and Power Sum

Suppose that the input sequence $s = (s_1, \ldots, s_L)$, where $s_i \in \{x_1, \ldots, x_k\}$ is generated from a given source $S$. The Shannon entropy is defined as

$$H(S) = H(p) = - \sum_{i=1}^{k} p_i \log_2 p_i,$$  

where $p = (p_1, \ldots, p_k)$ denotes the distribution of $S$.

The Rényi entropy of order $\alpha \in (0, 1) \cup (1, \infty)$ is defined as

$$H_\alpha(S) = H_\alpha(p) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{k} p_i^\alpha.$$  

For $\alpha = 2$, the Rényi entropy corresponds to the collision entropy $H_2(S)$ as follows:

$$H_2(S) = H_2(p) = - \log_2 \sum_{i=1}^{k} p_i^2.$$  

The min-entropy is defined as

$$H_\infty(S) = H_\infty(p) = - \log_2 \theta,$$  

where

$$\theta = \max_{i \in \{1, \ldots, k\}} \{p_i\}.$$  

Definition 1 (Power Sum): The power sum of order $\alpha$ (i.e., the $\alpha$th moment) for a distribution $p$ is defined as

$$M_\alpha(p) = \sum_{i=1}^{k} p_i^\alpha.$$  

Remark 2 (Collision Probability): The power sum of order $\alpha = 2$ (i.e., $M_2(p)$) is equivalent to the collision probability, which is the probability that two arbitrary source outputs are equal.

Remark 3: The Rényi entropy of order $\alpha$ is $H_\alpha(p) = \frac{1}{1-\alpha} \log_2 M_\alpha(p)$.

Remark 4: The following relations are well-known:

$$H(S) = \lim_{\alpha \to 1} H_\alpha(S),$$  

$$H_\infty(S) = \log_2 H_\alpha(S),$$  

Remark 5: The Rényi entropy is non-increasing in $\alpha$ [12]. Hence, $\forall \alpha, H_\infty(S) \leq H_\alpha(S)$, i.e., the min-entropy is a lower bound on the Shannon entropy and the Rényi entropy.

B. LRS Estimator and Its Overestimation Problem

For non-IID sources, NIST SP 800-90B proposes ten different min-entropy estimators (see Table I). These estimators independently perform their own estimations based on different statistics calculated from the examined non-IID sources.

| Statistic-based estimator [2] | Prediction-based estimator [3] |
|-------------------------------|-------------------------------|
| Most common value estimator | MultiMCW prediction estimator |
| Collision estimator | Lag prediction estimator |
| Markov estimator | MultiMM prediction estimator |
| Compression estimator | LZ78Y prediction estimator |
| $r$-Tuple estimator | $L$-RS estimator |

**Algorithm 1 LRS estimator of NIST 800-90B [1]**

**Input:** Sequence $s = (s_1, \ldots, s_L)$ where $s_i \in \{x_1, \ldots, x_k\}$.

**Output:** Collision entropy $H_2(S)$.

1. Find the smallest $u$ such that the number of occurrences of the most common $u$-tuple in $s$ is less than 35.
2. Find the largest $v$ such that the number of occurrences of the most common $v$-tuple in $s$ is at least 2. ▷ Longest repeated substring problem
3. for $w \in \{u, u+1, \ldots, v\}$ do
4.     Estimate the estimated $w$-tuple collision probability:
5.     $P_w := \frac{\sum_{i=1}^{L} \binom{C_i}{\binom{L}{2}}}{\binom{L}{2}}.$
6.     $\tilde{P}_w := P_w^{1/w}$.
7.     $\tilde{p}_c := \max \{ \tilde{P}_u, \ldots, \tilde{P}_v \}$.
8.     $\tilde{p}_c := \min \{ 1, \tilde{p}_c + 2.576 \sqrt{\frac{\tilde{p}_c (1-\tilde{p}_c)}{L-1}} \}$.
9.     $H_2(S) := - \log_2 \tilde{p}_c$.  

The min-entropy is defined as

$$H_\infty(S) = H_\infty(p) = - \log_2 \theta,$$  

where

$$\theta = \max_{i \in \{1, \ldots, k\}} \{p_i\}.$$  

Theoretical analysis and empirical results support that the generalized LRS estimator significantly improves the estimation accuracy of the LRS estimator, although both the LRS estimator and the proposed estimator rely on the same statistics, i.e., counts of repeated substrings. We believe that the generalized LRS estimator is an appealing alternative to the LRS estimator of NIST SP 800-90B since the generalized LRS estimator outputs a more accurate and stable min-entropy estimate from the same statistics of a given sequence.
Among these ten estimators, the $t$-tuple estimator and the LRS estimator compute entropies based on the frequency of substrings (tuples) in the input sequence $s$. The $t$-tuple estimator estimates the min-entropy based on the frequency of some fixed-length repeated substrings. The LRS estimator handles substring sizes that are too large for the $t$-tuple estimator [1], [11].

Algorithm 1 describes the LRS estimator in NIST SP 800-90B. Step 1 finds the smallest $u$ such that the number of occurrences of the most common $u$-tuple is less than 35. Step 2 solves the well-known longest repeated substring problem and sets $v$ as its length. Then, the range of $w$ becomes $\{u, u+1, \ldots, v\}$. In contrast, the $t$-tuple estimator finds the largest $t$ such that the number of occurrences of the most common $t$-tuple is at least 35 and the range of $w$ in that test equals $\{1, \ldots, t\}$ where $t < u$. Note that the $t$-tuple estimator and the LRS estimator calculate the entropies based on disjoint substring lengths, where the LRS estimator handles the longer substrings. Hence, the $t$-tuple estimator and the LRS estimator are complementary.

The LRS estimator estimates collision entropy instead of the min-entropy. Step 4 calculates the empirical collision probability of length-$w$ substrings. The LRS estimator of NIST SP 800-90B uses overlapped tuples, i.e., $l = L - w + 1$ in Step 4. For non-overlapped tuple counts, the total number of $w$-tuples becomes $l = \left\lfloor \frac{L}{w} \right\rfloor$.

The collision probability estimation by (9) is a key step of the LRS estimator, which was considered in [13], [14] for testing whether a distribution is close to uniform. Note that (9) is an unbiased estimator of the collision probability [15].

Step 5 computes the collision probability per sample (to normalize the estimated entropy) and Step 7 conservatively chooses the maximum (across $w$) collision probability (i.e., the minimum collision entropy). Step 8 ensures the confidence level of 99% under the Gaussian assumption, which is a common operation for all the min-entropy estimators of NIST SP 800-90B. Although Step 7 and Step 8 follow the conservative approach of NIST SP 800-90B, the LRS estimator overestimates the min-entropy since Step 9 estimates the collision entropy instead of the min-entropy.

Fig. 1 shows the ramifications of the fact that the LRS estimator estimates the collision entropy instead of the min-entropy. The bias between the actual min-entropy and the estimate by the LRS estimator is considerable except for $p = 0.5$. The numerical results in [16, Tables 2 and 3] also confirm this overestimation problem for the first-order Markov source and several pseudo-random data.

NIST SP 800-90B conservatively selects the minimum among estimated values by ten estimators. Hence, an overestimated value by the LRS estimator fails to exploit the important statistics of the longest repeated substring in a meaningful way, and may result in a security risk. Since the LRS estimator computes based on larger sizes of substrings than the $t$-tuple estimator, using the existing LRS estimator also results in significant waste of computational resources (its execution time is around ten times longer than that of the $t$-tuple estimator [11, Table VI]), which are expended regardless of the fact that the LRS estimator’s overestimated values rarely affects the final min-entropy estimate [1], [16, Table 2].

III. MIN-ENTROPY ESTIMATION BY LRS ESTIMATOR

A. Min-entropy Estimation by LRS Estimator

In this section, we propose a method to resolve the overestimation problem of the LRS estimator. The proposed method attempts to estimate the min-entropy instead of the collision entropy by using the estimation values of the LRS estimator and the following bound.

Lemma 6 ([9, Theorem 6]): Suppose that $\theta = \max_{i \in \{1, \ldots, k\}} \{p_i\}$. Then, the following inequality holds:

$$H_\alpha(S) \leq \frac{1}{1 - \alpha} \log_2 \left( \theta^\alpha + \frac{(1 - \theta)^\alpha}{(k - 1)\alpha - 1} \right)$$

(11)

for $\alpha \neq 1$. The bound is achieved with equality by the near-uniform distribution $p_{\text{NU}}(\theta) = (p_1, \ldots, p_k)$ where

$$p_i = \begin{cases} \theta, & \text{if } i = 1; \\ \frac{1 - \theta}{k - 1}, & \text{otherwise.} \end{cases}$$

(12)

Without loss of generality, $p_1 \geq \ldots \geq p_k$ is assumed.

The bound (11) is the counterpart of Fano’s inequality, which applies to the Shannon entropy.

Theorem 7: For the estimated collision probability $\hat{p}_c$ by Algorithm 1, the following inequality holds:

$$\theta \leq \sqrt{(k-1)(p_c(k-1) + 1)}$$

(13)

where $p_c = \mathbb{E}(\hat{p}_c)$. Since the near-uniform distribution achieves (11) with equality, (13) is the sharp upper bound.

Proof: For $\alpha > 1$, (11) leads to

$$M_\alpha(p) \geq \theta^\alpha + \frac{(1 - \theta)^\alpha}{(k - 1)\alpha - 1}.$$  

(14)

The term “sharp bound” means that there exists a distribution that achieves this bound with equality.
Algorithm 2 Proposed Estimator (Improved LRS Estimator)

Input: Sequence $s = (s_1, \ldots, s_L)$ where $s_i \in \{x_1, \ldots, x_k\}$.
Output: Min-entropy $H_\infty(S)$.

1: Estimate $\hat{p}_c$ from $s$ by Algorithm 1.
2: if $\hat{p}_c > \frac{1}{2}$ then
3: \hspace{1cm} $\hat{\theta} := \sqrt{(k-1)(\hat{p}_c k - 1)} + 1$.
4: else
5: \hspace{1cm} $\hat{\theta} := \frac{1}{k}$.
6: end if
7: $\hat{\theta} := \min \left(1, \hat{\theta} + 2.576 \sqrt{\frac{(1-\theta)}{\theta-1}}\right)$.
8: $H_\infty(S) := -\log_2 \hat{\theta}$.

Since $M_2(p)$ is equivalent to the collision probability $p_c$ [15], we can set $p_c \geq 3^2 + \frac{(1-\theta)^2}{(k-1)^2}$. By (5), it is clear that $\theta \geq \frac{1}{2}$. Since $\theta^2 + \frac{(1-\theta)^2}{(k-1)^2}$ is a non-decreasing function of $\theta$ for $\theta \geq \frac{1}{2}$, (13) holds.

Notice that (13) provides the optimal upper bound on $\theta$ since it is the sharp bound. Based on Theorem 7, we estimate $\hat{\theta}$ as follows:

$$\hat{\theta} = \frac{1}{k} \left(1 + \frac{1}{k} \sum_{i=1}^{k-1} \frac{1}{\psi^2} \right)$$

which is a conservative min-entropy estimation, i.e., $\theta \leq \hat{\theta}$ (compare (13) and (15)). Note that an upper bound on $\theta$ leads to a lower bound on $H_\infty(S)$, i.e.,

$$H_\infty(S) = -\log_2 \theta \geq -\log_2 \hat{\theta}$$

Algorithm 2 describes the proposed min-entropy estimator. Algorithm 2 enables estimating the min-entropy instead of the collision entropy by leveraging the LRS estimator. Step 1 of Algorithm 2 estimates the collision probability by using Algorithm 1. Theoretically, $p_c \geq \frac{1}{2}$ where the equality is achieved by the uniform distribution. If $\hat{p}_c < \frac{1}{k}$, then we know that it results from estimation errors. Hence, in this case we set $\hat{\theta} = \frac{1}{k}$, which comes from $\hat{p}_c = \frac{1}{k}$. Step 7 ensures the confidence level of 99% as in Step 8 of Algorithm 1 following the conservative approach of NIST SP 800-90B.

The proposed estimator attempts to estimate a lower bound on the min-entropy whereas the LRS estimator estimates an upper bound on the min-entropy (i.e., collision entropy). The proposed estimator matches the conservative approach of NIST SP 800-90B. It is worth mentioning that the collision estimator and the compression estimator of NIST SP 800-90B (Table I) also estimate lower bounds on the min-entropy by using the near-uniform distribution as in the proposed estimator. Hence, the proposed estimator is better aligned with other estimators of NIST SP 800-90B than the LRS estimator.

Importantly, the proposed estimator is unbiased for binary sources (i.e., it estimates the min-entropy itself instead of the lower bound since any binary distributions are near-uniform). In the next subsection, we further investigate the proposed estimator’s bias properties.

### B. Bias of Proposed Estimator

We investigate the biases of the conventional LRS estimator and the proposed estimator. For the analysis, we neglect the step for 99% confidence interval. Hence, $\hat{p}_c$ and $\hat{\theta}$ instead of $p_c$ and $\theta$ are considered in our analysis.

The bias is the difference between the estimator’s expected value and the true value. The bias of the LRS estimator is given by

$$b_{LRS} = E(-\log_2 \hat{p}_c) - H_\infty(S).$$

**Proposition 8:** The LRS estimator is overestimating, i.e., $b_{LRS} > 0$.

Proof: We show that $b_{LRS} > 0$ as follows:

$$b_{LRS} = E(-\log_2 \hat{p}_c) - H_\infty(S) > -\log_2 E(\hat{p}_c) - H_\infty(S) = H_2(S) - H_\infty(S) \geq 0,$$

where (18) follows from Jensen’s inequality. Since $-\log x$ is strictly convex and $\hat{p}_c$ is not constant for non-deterministic sources, (18) holds. Since $\hat{p}_c$ is obtained by an unbiased estimator for collision probability (i.e., $p_c = E(\hat{p}_c)$) [15], we claim that (19) holds. Also, (20) follows from Remark 5. Hence, $b_{LRS} > 0$.

As shown in Fig. 1, $H_2(S) - H_\infty(S)$ can be large for BMS with $p < 0.5$. Hence, the LRS estimator suffers from the severe overestimation problem.

The bias of the proposed estimator is given by

$$b_{proposed} = E(-\log_2 \hat{\theta}) - H_\infty(S) > -\log_2 \frac{\theta}{E(\hat{\theta})},$$

where $-\log_2 \frac{\theta}{E(\hat{\theta})} \leq 0$ since $\hat{\theta}$ is an estimate of the upper bound on $\theta$ (see Theorem 7). If $E(-\log_2 \hat{\theta}) \approx -\log_2 E(\hat{\theta})$, then the proposed estimator would be an underestimated estimator.

Since NIST SP 800-90B conservatively estimates the min-entropy, the proposed estimator is better aligned with NIST SP 800-90B than the LRS estimator. Further, we will show that the proposed estimator is unbiased for binary sources (see Corollary 14 and Remark 15).

We characterize the bias $b_{proposed}$ by the sharp lower and upper bounds on $\theta$ for a given collision probability $p_c$. The sharp upper bound on $\theta$ is given in Theorem 7. We derive the sharp lower bound on $\theta$ by using the inverted near-uniform distribution. In [2], the inverted near-uniform distribution is defined as $p_{INU}(\psi) = \{p_1, \ldots, p_k\}$ where

$$p_i = \begin{cases} \psi, & \text{if } i \in \left\{1, \ldots, \left\lfloor \frac{1}{\psi} \right\rfloor \right\}; \\ 1 - \left[ \frac{1}{\psi} \right] \psi, & \text{if } i = \left\lfloor \frac{1}{\psi} \right\rfloor + 1; \\ 0, & \text{otherwise}. \end{cases}$$

Note that $\psi = \max\{p_{INU}(\psi)\}$.

**Lemma 9:** For $\frac{1}{n+1} < \psi \leq \frac{1}{n}$ where $n \in \mathbb{N}$, the following relation holds:

$$\frac{1}{n+1} \leq \frac{1}{M_2(p_{INU}(\psi))} = n.$$
Min-entropy

Proof: If \( \psi = \frac{1}{n} \), then
\[
M_2(p_{\text{INU}}(\psi)) = \frac{1}{n}.
\]  
(24)

Hence, (23) holds.

If \( \frac{1}{n+1} < \psi < \frac{1}{n} \), then \( M_2(p_{\text{INU}}(\psi)) \) is an increasing function of \( \psi \). It is because 
\[
\frac{dM_2(p_{\text{INU}}(\psi))}{d\psi} = 2n(n+1)\{\psi - \frac{1}{n+1}\} > 0.
\]
By (24), we obtain \( \frac{1}{n+1} < M_2(p_{\text{INU}}) < \frac{1}{n} \). Then, (23) holds.

Remark 10: For an inverted near uniform distribution, Lemma 9 shows that the collision entropy is close to the min-entropy since \( \psi \approx M_2(p_{\text{INU}}) \). If \( \psi = \frac{1}{n} \), then the collision entropy is the same as the min-entropy.

Theorem 11: For any distribution \( p = (p_1, \ldots, p_k) \) with \( n = \left\lfloor \frac{1}{p_c} \right\rfloor \), the following inequalities hold:
\[
\psi \leq \theta \leq \hat{\theta},
\]
where
\[
\psi = \sqrt{n\{p_c(n+1)-1\} + n}, \quad \text{and} \quad \hat{\theta} = \frac{\sqrt{(k-1)(p_c k - 1)} + 1}{k}.
\]

Proof: Since \( \hat{\theta} \) is derived in Theorem 7, we need to derive only the sharp lower bound \( \psi \). The lower bound \( \psi \) is achieved with equality by the inverted near-uniform distribution \( p_{\text{NU}}(\psi) \) [2], [17]. Hence, we need to identify \( p_{\text{NU}}(\psi) \) satisfying \( M_2(p_{\text{NU}}(\psi)) = p_c \). Suppose that \( \frac{1}{n+1} < \psi < \frac{1}{n} \) where \( n \in \mathbb{N} \) (i.e., \( \left\lfloor \frac{1}{\psi} \right\rfloor = n \)). By (22) and Lemma 9, we obtain \( M_2(p_{\text{INU}}(\psi)) = n\psi^2 + (1-n\psi)^2 = p_c \), which leads to (26).

Remark 12: For a given collision probability \( p_c \), the sharp lower and upper bounds on the min-entropy are given by
\[
-\log_2 \theta \leq H_\infty(p) \leq -\log_2 \psi,
\]
where \( H_\infty(p) = -\log_2 \theta \).

We note that \( \psi \) depends only on \( p_c \) because \( n = \left\lfloor \frac{1}{p_c} \right\rfloor \). On the other hand, \( \hat{\theta} \) depends on \( p_c \) and \( k \) (alphabet size \( |S| \)). For given \( p_c \) and \( k \), we define the estimation gap of \( \theta \) as
\[
g(p_c, k) = \hat{\theta} - \psi,
\]
which is the maximum possible bias. The following theorem shows that the estimation gap increases with \( k \).

Theorem 13: For non-deterministic sources, the estimation gap \( g(p_c, k) = \hat{\theta} - \psi \) increases with \( k \).

Proof: Since \( \psi \) does not depend on \( k \), we show that \( \hat{\theta}(k) \) is an increasing function of \( k \). The derivative of \( \hat{\theta}(k) \) is given by
\[
\frac{d\hat{\theta}(k)}{dk} = \frac{p_c k + k - 2 - 2\sqrt{(k-1)(p_c k - 1)}}{2k^2\sqrt{(k-1)(p_c k - 1)}}.
\]
(30)

We can set \( p_c > \frac{1}{k} \) because \( p_c = \frac{1}{k} \) means that \( p \) is the uniform distribution, i.e., \( \psi = \hat{\theta} = \frac{1}{k} \) and \( H_\infty(p) = \log_2 k \). By the arithmetic-geometric mean inequality,
\[
p_c k + k - 2 = (p_c k - 1) + (k - 1) \geq 2\sqrt{(k - 1)(p_c k - 1)}.
\]
(31)

Hence, \( \frac{d\hat{\theta}(k)}{dk} > 0 \) for \( p_c < 1 \). Note that \( p_c < 1 \) for non-deterministic sources.

Corollary 14: For binary sources with \( k = 2 \), the estimation gap is zero, i.e., \( g(p_c, k = 2) = 0 \).

Proof: For binary sources, it is clear that \( \theta \geq \frac{1}{2} \). For \( \theta = \frac{1}{2} \), \( p \) corresponds to the binary uniform distribution, the gap is zero. For \( \theta > \frac{1}{2} \), we can set \( n = 1 \) in (23) because \( \frac{1}{2} < \psi < 1 \). By setting \( n = 1 \) and \( k = 2 \), (26) and (27) are identical, i.e., \( \psi = \hat{\theta} = \frac{\sqrt{2p_c - 1} + 1}{2} \).

Remark 15 (Unbiasedness): The proposed estimator is unbiased for binary sources. Since most random sources are binary or can be represented by binary sequences, the proposed estimator can effectively improve the bias of the LRS estimator.

Fig. 2(a) shows that the near-uniform distribution and the inverted near-uniform distribution correspond to the upper and lower bounds on \( \theta \), respectively. Also, it shows that \( \psi = \hat{\theta} \) for \( k = 2 \), i.e., the proposed estimator is unbiased for binary sources. Fig. 2(b) shows the relation between the collision entropy and the min-entropy where the near-uniform distribution and the inverted near-uniform distribution correspond to the lower and upper bounds on the min-entropy.
Algorithm 3 Generalized LRS estimator

**Input:** Sequence \( s = (s_1, \ldots, s_L) \) and an integer \( \alpha \geq 3 \)

**Output:** Min-entropy \( H_{\infty}(S) \).

1. Find the smallest \( u \) such that the number of occurrences of the most common \( w \)-tuple in \( s \) is less than 35.
2. Find the largest \( v \) such that the number of occurrences of the most common \( v \)-tuple in \( s \) is at least \( \alpha \).
3. for \( w \in \{ u, u+1, \ldots, v \} \) do
   4. Estimate the \( w \)-tuple power sum of order \( \alpha \):
      \[
      \hat{M}_{\alpha,w} := \sum \frac{C_i^w}{l^\alpha},
      \]  
      (32)
      where \( C_i \) is the number of occurrences of the \( i \)th unique \( w \)-tuple and \( l \) is the total number of \( w \)-tuples.
6. end for
7. \( \hat{M}_{\alpha} := \max \{ \hat{M}_{\alpha,u}, \ldots, \hat{M}_{\alpha,v} \} \).
8. if \( \hat{M}_{\alpha} > \frac{1}{\alpha^{w-1}} \) then
   9. By the bisection method, solve the following equation for \( \tilde{\theta} \in [\frac{1}{k}, 1] \):
      \[
      \hat{M}_{\alpha} = \tilde{\theta}^\alpha + \frac{(1 - \tilde{\theta})^\alpha}{(k - 1)^{\alpha-1}}.
      \]  
      (33)
   10. else
   11. \( \tilde{\theta} := \frac{1}{k} \).
12. end if
13. \( \hat{\theta} := \min \left( 1, \tilde{\theta} + 2.576 \sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{L-1}} \right) \).
14. \( H_{\infty}(S) := -\log_2 \hat{\theta} \).

We note that the final min-entropy estimation could not be perfectly unbiased because of the two steps of the original LRS estimator. First, Step 7 of Algorithm 1 selects the maximum collision probability among \( v-u+1 \) candidates, which lowers the min-entropy estimates. Also, Step 8 of Algorithm 1 (or Step 7 of Algorithm 2) reduces the min-entropy estimates to ensure the confidence level of 99\%. These steps result from the conservative approach of NIST SP 800-90B. These extra-confidence steps are also included in the other estimators of NIST SP 800-90B.

**IV. GENERALIZED LRS ESTIMATOR**

In this section, we propose a generalized LRS estimator by using the power sum of order \( \alpha \geq 2 \) instead of the collision probability (the power sum of order \( \alpha = 2 \)). We show that the generalized LRS estimator reduces the variance of estimates as the order \( \alpha \) increases beyond 2.

**A. Generalized LRS Estimator**

The generalized LRS estimator is based on 1) the generalized power sum \( M_{\alpha}(p) \) and 2) the proposed technique in Algorithm 2.

The generalized LRS estimator is described in Algorithm 3. First, it estimates the power sum \( M_{\alpha}(p) \) for a given \( \alpha \) by Steps 1–7. Step 2 of Algorithm 3 is modified to estimate \( \hat{M}_{\alpha}(p) \).

Step 4 estimates the \( w \)-tuple power sum of order \( \alpha \) by counting the \( \alpha \)-wise collisions. Step 5 computes the power sum of order \( \alpha \) per sample (to normalize the estimated min-entropy) and Step 7 conservatively chooses the maximum among estimated power sums of \( \alpha \), which is denoted by \( \hat{M}_{\alpha} \).

The estimation \( \hat{M}_{\alpha}(p) \) by (32) is a key step, which generalizes (9) in Algorithm 1. The estimation by (32) is unbiased [15], [18]. However, \( \hat{M}_{\alpha} \) in Step 7 is an overestimate of \( M_{\alpha}(p) \), which leads to an underestimate of the min-entropy. We maintain this conservative approach as in the LRS estimator of Algorithm 1.

Afterward, we estimate \( \hat{\theta} \) from \( \hat{M}_{\alpha} \) in Steps 8–12. First, we note that \( \hat{M}_{\alpha} \geq \frac{1}{\alpha^{w-1}} \) where the equality is achieved by the uniform distribution. Hence, we set \( \tilde{\theta} = \frac{1}{k} \) in Step 11 if \( \hat{M}_{\alpha} < \frac{1}{\alpha^{w-1}} \). If \( \hat{M}_{\alpha} > \frac{1}{\alpha^{w-1}} \), then Step 9 estimates \( \tilde{\theta} \) by using the bisection method. Step 13 ensures the confidence level of 99\% under the Gaussian assumption. Finally, Step 14 estimates the min-entropy from \( \tilde{\theta} \).

Unlike Algorithm 2, Step 9 of Algorithm 3 solves the non-closed-form equation by the bisection method. For a given iteration number of the bisection method, i.e., \( M \), the complexity to solve (33) corresponds to \( O(M) \). Note that \( M \) determines the numerical accuracy of the bisection method [7]. Hence, Algorithm 3 incurs the additional complexity of \( O(M) \).

The proposed Algorithm 3 improves the bias and reduces the variance compared to the LRS estimator (Algorithm 1). The bias is improved since Algorithm 3 estimates the min-entropy whereas the LRS estimator estimates the collision entropy as discussed in Section III-B. The variance can be reduced by using the higher-order power sum instead of the collision probability, which is supported by empirical results in Section V. In the following subsection, we provide a theoretical analysis of how the order \( \alpha \) affects the variance of estimation.

**B. Variance of Generalized LRS Estimator**

In this subsection, we attempt to characterize how the order \( \alpha \) affects the variance of \( \hat{\theta} \) calculated by (33) in Algorithm 3.

Suppose that \( \hat{\theta}_{\alpha} \) and \( \hat{\theta}_{\alpha+1} \) are the estimated \( \tilde{\theta} \) in Algorithm 3 by using \( \hat{M}_{\alpha} \) and \( \hat{M}_{\alpha+1} \), respectively. We characterize the relation between \( \alpha \) and the variance \( \text{Var}(\tilde{\theta}) = \text{E}(\tilde{\theta}^2) - \text{E}^2(\tilde{\theta}) \). We assume that the length-\( w \) tuples counted in Algorithm 3 are non-overlapping to simplify the analysis.

**Theorem 16:** For a uniformly distributed \( s = (s_1, \ldots, s_L) \) with a large \( L \), the variance ratio’s dependence on \( \alpha \) is as follows:

\[
\xi(\alpha) = \frac{\text{Var}(\hat{\theta}_{\alpha+1})}{\text{Var}(\hat{\theta}_{\alpha})} \approx \left( \frac{\alpha}{\alpha + 1} \right)^4,
\]  

(34)

where \( \approx \) hides multiplicative terms that tend to 1 as \( L \) goes to infinity.

**Proof:** The proof is given in Appendix B.

Since \( \xi(\alpha) < 1 \), \( \text{Var}(\tilde{\theta}) \) decreases with \( \alpha \) for high-entropy sources. The reduction of \( \text{Var}(\hat{\theta}_{\alpha}) \) diminishes as \( \alpha \) increases.

The range of \( w \in \{ u, u+1, \ldots, v \} \) is an important parameter that affects the variance of \( \tilde{\theta} \). It is clear that

\[
v_u \geq v_{u+1}, \quad u = u_u = u_{\alpha+1},
\]  

(35)
where $\nu_\alpha$ and $\nu_{\alpha+1}$ are calculated by Step 2 of Algorithm 3 for $\alpha$ and $\alpha+1$, respectively. Note that $u$ in Algorithm 3 does not depend on $\alpha$. The proof of Theorem 16 relies on this relation since the reduction of $v$ leads to the reduction of $\text{Var}(\hat{\theta})$.

Theorem 16 characterizes $\text{Var}(\hat{\theta})$ by (33), i.e., for $\tilde{M}_\alpha > \frac{1}{k^{\alpha-1}}$. If $\tilde{M}_\alpha < \frac{1}{k^{\alpha-1}}$, then Step 11 sets $\hat{\theta} := \frac{1}{k}$. It is because the power sum of order $\alpha$ cannot be lower than $\frac{1}{k^{\alpha-1}}$ (attained by the uniform distribution). It is difficult to analyze the probability of $\tilde{M}_\alpha < \frac{1}{k^{\alpha-1}}$ due to Step 7 of $\tilde{M}_\alpha := \max\{\tilde{M}_{\alpha,1}, \ldots, \tilde{M}_{\alpha,v}\}$ in Algorithm 3.

Although Theorem 16 focuses on uniformly distributed sources, the following section empirically supports that $\text{Var}(\hat{\theta})$ decreases with $\alpha$ even for non-uniformly distributed sources.

Remark 17: We leverage the Rényi entropy to reduce the variance of min-entropy estimation in Algorithm 3. In [7], the Rényi entropy is exploited to resolve the underestimation problem of the compression estimator. It is worth mentioning that the compression estimator suffers from the underestimation problem whereas the LRS estimator suffers from the overestimation problem.

V. NUMERICAL RESULTS

We evaluate our proposed estimators for simulated and real-world data samples. The empirical results show that the proposed estimator effectively reduces the bias problem of the LRS estimator.

The following representative samples are considered as in [3], [7]:

- **Binary memoryless source (BMS):** Samples are generated by Bernoulli distribution with $P(S = 1) = p$ and $P(S = 0) = 1 - p$ (IID);
- **Markov source:** Samples are generated using the first-order Markov model with $P(S_{t+1} = 1|S_t = 0) = P(S_{t+1} = 0|S_t = 1) = p$ (non-IID);
- **Near-uniform distribution:** Samples are generated by the near-uniform distribution with $k = 64$ (see (12)) (IID);
- **Inverted near-uniform distribution:** Samples are generated by the inverted near-uniform distribution with $k = 64$ (see (22)) (IID);
- **Normal distribution rounded to integers:** Samples are drawn from a normal distribution and rounded to integer values (IID).
- **Time-varying normal distribution rounded to integers:** Samples are drawn from normal distributions and rounded to integer values, where the mean values of the normal distributions are time-varying (non-IID).

For each of the above sources, one thousand simulated sources were created in each of the above datasets. BMS source and Markov source generate a sequence of $L = 100,000$ bits. The other sources generate a sequence of $L = 10,000$ bits and $k = 64$.

Fig. 3 compares the min-entropy estimators for BMS as a function of $p$. The theoretical min-entropy and collision entropy are given by $H_\infty(S) = -\log_2 \max\{p, 1-p\}$ and $H_2(S) = -\log_2\left\{p^2 + (1 - p)^2\right\}$, respectively. As discussed in Section II (viz. Fig. 1), the LRS estimator estimates the collision entropy instead of the min-entropy. Since $H_\infty(S) \leq H_2(S)$, the LRS estimator undesirably overestimates the min-entropy. For $p = 0.3$, the bias of the LRS estimator is around 0.28.

The proposed estimator accurately estimates the min-entropy as shown in Fig. 3(a). As $p \to 0.5$ (i.e., uniformly distributed sources), the higher $\alpha$ reduces $\text{Var}(\hat{\theta})$, which supports Theorem 16. We note that the reduction of $\text{Var}(\hat{\theta})$ diminishes as $\alpha$ increases as shown in Fig. 3(b).

We observe in Fig. 3(a) that the higher $\alpha$ slightly improves the bias as $p \to 0.5$. It is surprising because (32) is an unbiased estimator for any $\alpha$. The bias improvement results from Step 11 of Algorithm 3. Since the power sum of order $\alpha$ cannot be lower than $\frac{1}{k^{\alpha-1}}$ for binary sources (i.e., $k = 2$), we set $\hat{\theta} = \frac{1}{k}$ for $\tilde{M}_\alpha < \frac{1}{k^{\alpha-1}}$, which leads to $H_\infty(S) = 1$. For a BMS source with $p = \frac{1}{2}$, the estimated min entropy would be 1 (for $\tilde{M}_\alpha \leq \frac{1}{k^{\alpha-1}}$), which happens more frequently for higher $\alpha$ values. Thus, the increase of $\alpha$ can simultaneously reduce $\text{Var}(\hat{\theta})$ and improve the bias for high-entropy sources.

It is worth mentioning that because of the finite sample size $L$, the higher order $\alpha$ reduces the number of valid $C_i$ due to the
Min-entropy

Fig. 4. (a) Estimated min-entropy and (b) the variance of min-entropy by the proposed generalized LRS estimator for the first-order Markov sources with $p = p(1|0) = p(0|1)$. The proposed estimator ($\alpha = 2$) corresponds to Algorithm 2, and the proposed estimators ($\alpha \geq 3$) correspond to Algorithm 3.

Fig. 5. Comparison of min-entropy estimators for (a) near-uniform distributed sources and (b) inverted near-uniform distributed sources. The proposed estimator ($\alpha = 2$) corresponds to Algorithm 2, and the proposed estimators ($\alpha \geq 3$) correspond to Algorithm 3.

The sequences generated by these non-binary sources are represented by binary values. Then, we estimate the min-entropies from these binary sequences.

Fig. 5(a) shows that the LRS estimator undesirably overestimates the min-entropy for near-uniform distributed sources since the gap between the collision entropy and the min-entropy is significant. On the other hand, the proposed estimators estimate the min-entropy accurately. We note that the generalization LRS estimator with $\alpha > 2$ (Algorithm 3) is more accurate than the improved LRS estimator (Algorithm 2).

For the inverted near-uniform distributed sources, the LRS estimator is relatively accurate (but still overestimating) since the collision entropy is close to the min-entropy as discussed in Remark 10. Fig. 5(b) also supports Remark 10. We observe that the LRS estimator is close to the collision entropy (and the min-entropy) although it slightly overestimates the only the variance of estimates but also the bias as $p \to 0.5$.

Note that the improvement of $\text{Var}(\hat{\theta})$ diminishes as $\alpha$ increases as shown in Fig. 4(b).

For the first-order Markov sources, the min-entropy estimators estimate the min-entropy rate. By [19], [20], the accurate min-entropy rate and the collision entropy rate are given by $H_\infty(S) = -\log_2 \max\{p, 1-p\}$ and $H_2(S) = -\log_2 \{p^2 + (1-p)^2\}$, respectively. Note that the entropy rates of the first-order Markov sources are the same as the entropies of the BMS.

Fig. 4 compares the min-entropy estimators for the first-order Markov sources with parameter $p$. The LRS estimator of NIST SP 800-90B undesirably overestimates the min-entropy of the Markov sources as shown in Fig. 4(a). The proposed estimator effectively improves the accuracy of min-entropy estimates. As in Fig. 3, the generalized estimator improves not requirement of $C_i \geq \alpha$ in (32). Then, (32) could underestimate the power sum of $\alpha$. Given the sample size $L$, an $\alpha$ value should be picked so as to satisfy $C_i \geq \alpha$ for values of $w$ as large as of interest to the randomness tester. In our experiments with $L = 100,000$, we recommend $\alpha \in \{3, 4, 5, 6\}$ by taking into account valid $C_i$ and diminishing variance reduction of $\alpha$.
min-entropy. The proposed estimators underestimate the min-entropy, which can be explained by (15). Since \( \hat{\theta} > \hat{\rho}_e \) in (15), we observe that \( \hat{\theta} > \hat{\rho}_e \simeq \theta \) for inverted near-uniform distributed sources. Fortunately, the underestimation bias can be reduced by adopting a higher order \( \alpha \) as shown in Fig. 5(b).

It is worth mentioning that the compression estimator of NIST SP 800-90B also suffers from this underestimation problem for inverted near-uniform distributed sources [1], [7]. Compared to the compression estimator, our proposed estimators are much more accurate. For an inverted near-uniform distributed source with \( H_\infty(S) = 5 \), the estimated value by the proposed estimator with \( \alpha = 6 \) is around 4.3, which is much better than the compression estimator’s value of 1.55 (see [7, Fig. 6(b)]). Since NIST SP 800-90B conservatively selects the minimum estimates among ten different estimators including the compression estimator, the proposed generalized LRS estimator does not degrade the final estimation accuracy even for this exceptional inverted near-uniform distribution.

Fig. 6 compares the min-entropy estimators for the sequences sampled from normal distributions (rounded to integers) and the sequences sampled from normal distributions (rounded to integers) whose mean values are time-varying.

Fig. 6(a) shows that the LRS estimator overestimates the min-entropy for normal distributions. The proposed estimators resolve this overestimation problem effectively. We observe that the proposed estimators underestimate the min-entropy of the normally distributed sources as in the inverted near-uniform distributed sources. Nonetheless, comparing Fig. 6(a) and [7, Fig. 7], the proposed estimator yields higher estimated min-entropy values than the compression estimator (another standard min-entropy estimator of NIST SP 800-90B). Thus, the negative impact of the proposed estimator’s underestimation would be restricted. While the proposed tests introduce a small degree of underestimation, in practice this can be mitigated by increasing the parameter \( \alpha \) when (and only when) the proposed test returns the lowest min-entropy estimate among all tests.

For the time-varying normal distribution, the mean value of the normal distribution jumps in the middle of the data sequence. The actual min-entropy is calculated by the average of two parts of the sequence as in [11]. Fig. 6(b) shows that the LRS estimator undesirably overestimates the min-entropy. The proposed estimators effectively improve the estimation accuracy, especially for \( \alpha = 6 \).

We also evaluate min-entropy estimates using random number generators deployed in the real world as in [3], [7]. The true min-entropies for these sources are unknown even though they are believed to be high-entropy sources. We evaluate RANDOM.ORG, Ubld.it, and Linux kernel random number generator (LKRNG). RANDOM.ORG [21] is a service that provides random numbers based on atmospheric noise and Ubld.it generates random numbers by a TrueRNG device by [22]. The min-entropy estimates of the real-world sources are presented in Table II. Given the numerical results shown earlier in the section, it is safe to assume that the LRS estimator overestimates the min-entropy of these sources, and the results in the table show that the proposed estimators with increasing \( \alpha \) values likely approach the true value from below.

The numerical results show that the LRS estimator suffers from significant overestimation for most sources. Hence, the proposed estimator would be an appealing alternative to the original LRS estimator.

**VI. Conclusion**

We proposed accurate min-entropy estimators to resolve the overestimation problem of the LRS estimator. Although the proposed estimator relies on the estimated collision probability as in the LRS estimator, it effectively reduces the bias by leveraging the relation between the collision entropy and the min-entropy. Furthermore, we proposed the generalized LRS

---

**TABLE II**

| Source       | LRS  | \( \alpha = 2 \) | \( \alpha = 4 \) | \( \alpha = 6 \) |
|--------------|------|-----------------|-----------------|-----------------|
| RANDOM.ORG   | 0.9836 | 0.8889          | 0.9549          | 0.9572          |
| Ubld.it      | 0.9347 | 0.8277          | 0.8598          | 0.8941          |
| LKRNG        | 0.9967 | 0.9364          | 0.9843          | 0.9844          |

---

**Fig. 6.** Comparison of min-entropy estimators for (a) normally distributed sources (b) time-varying normally distributed sources. The proposed estimator \((\alpha = 2)\) corresponds to Algorithm 2, and the proposed estimators \((\alpha \geq 3)\) correspond to Algorithm 3.
estimator by parameterizing $\alpha$ instead of setting $\alpha = 2$. It was shown that the generalized LRS estimator can improve the bias and variance of min-entropy estimates. Theoretical analysis on the estimation variance for non-uniformly distributed sources would be an important future work.

**APPENDIX A**

**ANALYSIS ON $v$**

In this appendix, we analyze $v$ (i.e., the maximum value of $w$), which is used in Algorithms 1 and 3. We denote the number of $\alpha$-wise collisions as $D_{\alpha,w}$ for the $w$-tuples in Step 4 of Algorithm 3, which is given by

$$D_{\alpha,w} = \sum_{i=1}^{k} \left( \begin{array}{c} C_i \\ \alpha \end{array} \right),$$

(36)

where $C_i$ is the number of occurrences of the $i$th $w$-tuple. We suppose that $\left( \begin{array}{c} C_i \\ \alpha \end{array} \right) = 0$ if $C_i < \alpha$. Note that (36) is the same as the numerator of (32).

The following lemma shows the relation between the tuple size $w$ and the number of $\alpha$-wise collisions (36).

**Lemma 18:** For a large $L$, $E(D_{\alpha,w+1}) \approx M_\alpha(p) \cdot E(D_{\alpha,w})$.

**Proof:** Denote $\{a_1, \ldots, a_k\}$ as the alphabet of $w$-tuples and $\{c_1, \ldots, c_k\}$ as the number of occurrences of each $w$-tuple in $s$. For a $w$-tuple element $a_j$ for $j \in \{1, \ldots, k\}$, we can represent $a_j$ as $(a_{j,1}, \ldots, a_{j,w})$ where $a_{j,i} \in \{x_1, \ldots, x_k\}$ for $i \in \{1, \ldots, w\}$.

For each $w$-tuple $a_j = (a_{j,1}, \ldots, a_{j,w})$, there are $k$ different ways to add a symbol and obtain a $(w+1)$-tuple. The expected numbers of occurrences with $a_j$ as prefix are $\{c_j \cdot p_1, c_j \cdot p_2, \ldots, c_j \cdot p_k\}$. Hence, the expected number of $\alpha$-wise collisions for $(w+1)$-tuples is given by

$$E(D_{\alpha,w+1}) = \sum_{j=1}^{k} \sum_{i=1}^{k} \left( \begin{array}{c} c_j \cdot p_i \\ \alpha \end{array} \right)$$

$$\approx \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{c_j \cdot p_i}{\alpha}! \approx \sum_{j=1}^{k} \frac{c_j}{\alpha}! \approx M_\alpha(p) \cdot \left( \begin{array}{c} C_j \\ \alpha \end{array} \right) \cdot M_\alpha(p)$$

$$= M_\alpha(p) \cdot E(D_{\alpha,w}),$$

(38)

where (37) follows from $\left( \begin{array}{c} c_j \cdot p_i \\ \alpha \end{array} \right) \approx \frac{c_j \cdot p_i^{\alpha}}{\alpha!}$ for $c_j \cdot p_i \gg \alpha$ (i.e., for a large $L$) and Definition 1. If $c_j \cdot p_i$ is not much greater than $\alpha$, then it can be neglected. Also, (38) follows from $\left( \begin{array}{c} C_j \\ \alpha \end{array} \right) \approx \frac{c_j}{\alpha}!$.

Finally, (39) follows from (38).

For the proof of Theorem 16, we will take the value of $v$ to be $\tau$, which is defined to be the tuple length at which the distribution attains in expectation the cutoff property of having at least one tuple occurring at least $\alpha$ times in the sequence (see Step 2 in Algorithm 3).

**Lemma 19:** For a large $L$, $\tau \approx \log_{\frac{1}{\alpha \ln 2}} \left( \frac{1}{\alpha} \right)$.

**Proof:** By the definition of $\tau$ and (36), $\tau$ is the largest value such that $E(D_{\alpha,\tau}) \geq 1$. Hence, $E(D_{\alpha,1}) \cdot (M_\alpha)^{\tau-1} \geq 1$ and $E(D_{\alpha,1}) \cdot (M_\alpha)^{\tau} < 1$ by Lemma 18. Then, we can obtain

$$\log_{\frac{1}{\alpha \ln 2}} \left( \frac{1}{\alpha} \right) - 1 < \tau \leq \log_{\frac{1}{\alpha \ln 2}} \left( \frac{1}{\alpha} \right).$$

(41)

For $l \gg \alpha$, $\tau \approx \log_{\frac{1}{\alpha \ln 2}} \left( \frac{1}{\alpha} \right)$. $\blacksquare$

**Lemma 20:** For a uniformly distributed $s = (s_1, \ldots, s_L)$ with a large $L$,

$$\frac{\tau_{\alpha+1}}{\tau_\alpha} \approx \frac{\alpha^2 - 1}{\alpha^2},$$

(42)

which is less than one. Note that $\tau_\alpha$ denotes the $\tau$ with order $\alpha$.

**Proof:** We can obtain

$$\frac{\tau_{\alpha+1}}{\tau_\alpha} = \frac{\log_{\frac{1}{\alpha \ln 2}} \left( \frac{1}{\alpha} \right)}{\log_{\frac{1}{\alpha \ln 2}} \left( \frac{1}{\alpha+1} \right)}$$

$$= \frac{\alpha - 1}{\alpha} \frac{\ln \left( \frac{1}{\alpha+1} \right)}{\ln \left( \frac{1}{\alpha} \right)}$$

$$\approx \frac{\alpha - 1}{\alpha} \frac{\ln L^\alpha - \ln \tau_\alpha - \ln (\alpha + 1)!}{\ln L^\alpha - \ln \tau_\alpha - \ln \alpha!}$$

$$\approx \frac{\alpha - 1}{\alpha} \frac{\ln L^\alpha - \ln \tau_\alpha - \ln \alpha!}{\ln L^\alpha - \ln \tau_\alpha - \ln \alpha!}$$

$$= \frac{(\alpha - 1)(\alpha + 1)}{\alpha^2},$$

(43)

where (43) follows from Lemma 19 and $M_\alpha = k^{-\alpha-1}$ for a uniformly distributed source. For a large $L$, (45) follows from $\left( \begin{array}{c} \alpha \end{array} \right) \approx \frac{\alpha}{e \alpha}!$ and $l_\alpha = \left\lfloor \frac{\alpha}{e \alpha} \right\rfloor \approx \frac{\alpha}{e \alpha}$. Also, (46) follows from $L^\alpha \gg \tau_\alpha^3$ and $L^\alpha \gg \alpha$ for a large $L$. $\blacksquare$

**APPENDIX B**

**PROOF OF THEOREM 16**

For every subset $I \subseteq \{1, \ldots, l = \lfloor \frac{p_i}{\alpha} \rfloor \}$ of size $\alpha$, we define $X_I$ to be a 0-1 random variable that gets the value 1 iff all the values $x_l$ are the same (i.e., $I$ forms an $\alpha$-wise collision). By (36), it is clear that

$$D_{\alpha,w} = \sum_{|I|=\alpha} X_I$$

and

$$E(X_I) = M_\alpha(w),$$

(48)

where $M_\alpha(w)$ is the $w$-tuple power sum of order $\alpha$. Also, we set $X_I = X_I - M_\alpha(w)$ as in [18].

For two subsets $I$ and $J$ such that $|I| = |J| = \alpha$, $E(X_I \cdot X_J) = E(X_I) \cdot E(X_J) = 0$ if $I \cap J = \emptyset$. If $I \cap J \neq \emptyset$, then $X_I \cdot X_J$ is a 0-1 random variable that gets the value 1 iff all the values in $I \cup J$ are the same. Hence,

$$E(X_I \cdot X_J) = M_{\alpha+t,w} - M^2_{\alpha,w},$$

(50)

where $M_{\alpha+w,w}$ is the $w$-tuple power sum of order $\alpha+t$. Also, we set $X_I \cdot X_J = X_I - M_\alpha(w)$ as in [18].

For two subsets $I$ and $J$ such that $|I| = |J| = \alpha$, $E(X_I \cdot X_J) = E(X_I) \cdot E(X_J) = 0$ if $I \cap J = \emptyset$. If $I \cap J \neq \emptyset$, then $X_I \cdot X_J$ is a 0-1 random variable that gets the value 1 iff all the values in $I \cup J$ are the same. Hence,

$$E(X_I \cdot X_J) = \frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}}.$$

(51)
The variance of $D_{\alpha,w}$ is given by
\[
\text{Var}(D_{\alpha,w}) = \sum_{t=0}^{\alpha-1} \sum_{|\cup \cup|=\alpha+t} \mathbb{E}(X_I \cdot X_J) = \sum_{t=0}^{\alpha-1} \sum_{|\cup \cup|=\alpha+t} \left( \frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}} \right) \quad (52)
\]
(53)
\[
= \sum_{t=0}^{\alpha-1} \left( \frac{1}{\alpha} \right) \left( \left( l - \alpha \right) \frac{\alpha}{t} \right) \left( \frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}} \right) \quad (54)
\]
\[
\approx \frac{1}{k^{w(\alpha-1)}} \sum_{t=0}^{\alpha-1} \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( \frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{w(\alpha-1)}} \right) \quad (55)
\]
where (53) follows from (51). Also, (55) follows from $\binom{l-\alpha}{t} \approx \frac{l}{\alpha} \binom{\alpha}{t}$ for $l \gg \alpha$ and $\frac{1}{\alpha} = \frac{k^{w(\alpha-1)}}{k^{w(\alpha-1)}} = 0$ for $t = \alpha - 1$.

By taking into account normalization in Step 5 of Algorithm 3, we obtain
\[
\text{Var}(\tilde{M}_{\alpha,w}) = \text{Var}(\tilde{M}_{\alpha,w}^{1/2}) \approx \frac{1}{w^2} \cdot \mathbb{E}(\tilde{M}_{\alpha,w})^{2(\alpha-1)} \cdot \text{Var}(\tilde{M}_{\alpha,w}) \quad (56)
\]
\[
= \frac{1}{w^2} \cdot k^{2(\alpha-1)(w-1)} \cdot \text{Var}(D_{\alpha,w}) \quad (57)
\]
\[
\approx \frac{k^{(\alpha-1)(w-2)}}{w^2} \cdot \sum_{t=0}^{\alpha-1} \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( k^{w-1} - k^{w(\alpha-1)} \right) \quad (58)
\]
where (56) follows from the first-order Taylor approximation (i.e., $\text{Var}(f(x)) \approx f'(\mathbb{E}(x))^2 \cdot \text{Var}(x)$ where $f(x) = x^{1+\frac{1}{k}}$).

Since (32) is an unbiased estimator (i.e., $\mathbb{E}(\tilde{M}_{\alpha,w}) = M_{\alpha,w} = k^{-w(\alpha-1)}$), (57) holds. Finally, (58) follows from (55).

Now we show that $\text{Var}(\tilde{M}_{\alpha,w}) \leq \text{Var}(\tilde{M}_{\alpha,w+1})$ for $w \geq 3$.

For each term of (58),
\[
\frac{k^{(\alpha-1)(w-2)}}{w^2} \cdot \sum_{t=0}^{\alpha-1} \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( k^{w-1} - k^{w(\alpha-1)} \right) < \frac{k^{(\alpha-1)(w-2)}}{(w+1)^2} \cdot \frac{k^{\alpha-1}}{(\alpha)} \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( k^{w-1} - k^{w(\alpha-1)} \right) \quad (59)
\]
\[
\approx \frac{k^{(\alpha-1)(w-1)}}{(w+1)^2} \cdot \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( k^{w+1-t} - k^{w(\alpha-1)+t} \right) \quad (60)
\]
\[
< \frac{k^{(\alpha-1)(w-1)}}{(w+1)^2} \cdot \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( k^{w+1-t} - k^{w(\alpha-1)+t} \right) \quad (61)
\]
where (59) follows from $\binom{w}{w+1}^2 \cdot \frac{k^{\alpha-1}}{k^w} \geq 1$ for $k \geq 2$ and $w \geq 3$. Also, (61) follows from $t < \alpha - 1$. Hence,
\[
\text{Var}(\tilde{M}_{\alpha,w}) < \text{Var}(\tilde{M}_{\alpha,w+1}) \quad (62)
\]
for $w \geq 3$.

As a conservative approach, we set $\text{Var}(\tilde{M}_{\alpha}) \approx \text{Var}(\tilde{M}_{\alpha,w})$.

Then,
\[
\text{Var}(\tilde{M}_{\alpha}) \approx \frac{k^{(\alpha-1)(\alpha-2)}}{\pi_{\alpha}^2} \cdot \sum_{t=0}^{\alpha-2} \left( \frac{l}{\alpha} \right) \left( \frac{\alpha}{t} \right) \left( k^{t \cdot \pi_{\alpha} - k^{(\alpha-1) \pi_{\alpha}}-1} \right) \quad (63)
\]
where we denote $\pi = \pi_{\alpha}$ and $l = l_{\alpha}$ since both $\pi$ and $l$ depend on $\alpha$.

By the first-order Taylor approximation,
\[
\text{Var}(\tilde{\theta}_{\alpha}) \approx z(\tilde{\theta}_{\alpha}, \alpha)^2 \cdot \text{Var}(\tilde{M}_{\alpha}), \quad (64)
\]
where $z(\tilde{\theta}_{\alpha}, \alpha)$ is given by
\[
\frac{d \tilde{\theta}_{\alpha}}{d M_{\alpha}} = \frac{1}{\alpha} \left( \tilde{\theta}_{\alpha} - 1 - \frac{\tilde{\theta}_{\alpha}}{k \cdot \pi_{\alpha}} \right) \quad (65)
\]
which is derived from (33).

If $\tilde{M}_{\alpha} = \frac{1}{\pi_{\alpha}}$, then $\tilde{\theta}_{\alpha} \rightarrow \infty$. However, Algorithm 3 sets $\tilde{\theta}_{\alpha} = \frac{1}{k} \tilde{M}_{\alpha} \leq \frac{1}{k} \pi_{\alpha}$ instead of solving (33). Hence, $\tilde{\theta}_{\alpha}$ should be considered only if $M_{\alpha} = \frac{1}{k} \pi_{\alpha} + \delta$ where $0 < \delta \ll k$ for uniformly distributed sources. Then, we can set $\tilde{\theta}_{\alpha} = \frac{1}{k} + \delta'$ where $0 < \delta' \ll k$. By [7, Theorem 4], $\tilde{\theta}_{\alpha} \approx \frac{k^{\alpha-3}}{\alpha \cdot (\alpha-1)} \cdot \frac{1}{k \cdot \delta}$. Then,
\[
\frac{z(\tilde{\theta}_{\alpha+1}, \alpha+1)}{z(\tilde{\theta}_{\alpha}, \alpha)} \approx \frac{\alpha - 1}{\alpha+1} \cdot k \quad (66)
\]
Then, we obtain
\[
\xi(\alpha) = \frac{\text{Var}(\tilde{\theta}_{\alpha+1})}{\text{Var}(\tilde{\theta}_{\alpha})} \approx \frac{z(\tilde{\theta}_{\alpha+1}, \alpha+1)^2}{z(\tilde{\theta}_{\alpha}, \alpha)^2} \cdot \text{Var}(\tilde{M}_{\alpha+1}) \quad (67)
\]
\[
= \frac{(\alpha - 1)^2}{\alpha + 1} \cdot \frac{\pi_{\alpha}}{\alpha+1} \cdot \frac{\binom{\alpha}{\alpha+1}}{\binom{\alpha}{\alpha+1}} \cdot k^{(\alpha \pi_{\alpha+1} - (\alpha-1) \pi_{\alpha})} \cdot \sum_{t=0}^{\alpha-1} \binom{\alpha}{\alpha+1} \left( \frac{\alpha}{t} \right) \left( k^{t \cdot \pi_{\alpha} - k^{(\alpha-1) \pi_{\alpha}}} \right) \quad (68)
\]
\[
\approx \frac{(\alpha - 1)^2}{\alpha + 1} \cdot \frac{\pi_{\alpha}}{\alpha+1} \cdot \frac{\binom{\alpha}{\alpha+1}}{\binom{\alpha}{\alpha+1}} \cdot \sum_{t=0}^{\alpha-1} \left( \frac{\alpha}{t} \right) \left( k^{t \cdot \pi_{\alpha} - k^{(\alpha-1) \pi_{\alpha}}} \right) \quad (69)
\]
where (67) follows from (64). Also, (68) follows from (63) and (66). By Lemma 19 (see Appendix A) and $M_{\alpha} = \frac{1}{k \cdot \pi_{\alpha}}$, we obtain $\pi_{\alpha} \approx \log_{k^{\alpha-1}} \left( \frac{\alpha}{\alpha} \right) = \frac{1}{\alpha-1} \log_k \left( \frac{\alpha}{\alpha} \right)$, which leads to
\[
k^{\pi_{\alpha}} \approx \frac{\alpha}{\alpha} \quad (70)
\]
Then, (69) follows from $k^{(\alpha \pi_{\alpha+1} - (\alpha-1) \pi_{\alpha})} = \frac{\alpha}{\alpha}$.
Also, we obtain
\[
\xi(\alpha) \approx \left(\frac{\alpha}{\alpha + 1}\right)^4 \sum_{t=0}^{\alpha-1} \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{71}
\]
\[
= \left(\frac{\alpha}{\alpha + 1}\right)^4 \sum_{t=0}^{\alpha-2} \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{72}
\]
\[
\approx \left(\frac{\alpha}{\alpha + 1}\right)^4 \sum_{t=0}^{\alpha-1} \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}. \tag{73}
\]
where (71) and (72) follow from Lemma 20 and (70), respectively. Also, (73) follows from
\[
\left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1} \gg \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{74}
\]
for \(t \leq \alpha - 2\) and a large \(L\). We derive (74) as follows:
\[
\left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1} \geq \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{75}
\]
\[
> \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{76}
\]
\[
\geq \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right), \tag{77}
\]
where (75) follows from \(t \leq \alpha - 2\) and (76) follows from \(\left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \gg \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \frac{2}{\alpha + 1}\) for \(l_{t+1} > \alpha\). Also, (77) holds because \(l_{t+1} \gg \alpha\).

Finally, we show that (73) converges to \(\left(\frac{\alpha}{\alpha + 1}\right)^4\). For a large \(L, l_{t+1} = \frac{l_{t+2}}{l_t} \approx \frac{l_{t+2}}{l_t} \approx \frac{l_{t+2}}{l_t} \). Then,
\[
\sum_{t=0}^{\alpha-1} \left(l_{t+1} \right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{78}
\]
\[
= \sum_{t=0}^{\alpha-1} \left(l_{t+1} \right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}, \tag{79}
\]
\[
= \sum_{t=0}^{\alpha-1} \left(l_{t+1} \right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}. \tag{80}
\]
Also,
\[
\sum_{t=0}^{\alpha-2} \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1} \approx \sum_{t=0}^{\alpha-2} \left(\frac{l_{t+1}}{l_t}\right) \left(\frac{\alpha + 1}{t} \right) \left(\frac{l_{t+1} - \alpha \sigma_{t+1}}{l_t} \right) \frac{2}{\alpha + 1}. \tag{81}
\]
For a large \(L\), (80) converges to one because the highest degree of \(L^{-\frac{t}{l}}\) is zero by \(t = 0\). Similarly, (81) converges to one. Hence, \(\xi(\alpha) \approx \left(\frac{\alpha}{\alpha + 1}\right)^4\) for a large \(L\).
Jiheon Woo (Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and computer science from the Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, in 2021 and 2023, respectively. He is currently pursuing the Ph.D. degree in electrical engineering, Pohang University of Science and Technology (POSTECH), Pohang, South Korea. His research interests include information theory and machine learning.

Chanhee Yoo is currently pursuing the B.S. degree with the Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea. He was a Research Intern in the Department of Electrical Engineering and Computer Science at DGIST. His research interests include information theory, cryptography, and computational geometry.

Young-Sik Kim (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University in 2001, 2003, and 2007, respectively. He joined the Semiconductor Division of Samsung Electronics, where he worked in the research and development of security hardware IPs for various embedded systems, including modular exponentiation hardware accelerator (called Tornado 2MX2) for RSA and elliptic-curve cryptography in smart-card products and mobile application processors of Samsung Electronics, until 2010. He is currently a Professor at Chosun University, Gwangju, South Korea. He is also aSubmitter for two candidate algorithms (McNie and pqsigRM) in the first round for the NIST Post Quantum Cryptography Standardization. His research interests include post-quantum cryptography, fully homomorphic encryption, privacy-preserving machine learning, channel coding, and signal design. He was selected as one of 2025’s 100 Best Technology Leaders (for Crypto-Systems) by the National Academy of Engineering of Korea.

Yuval Cassuto (Senior Member, IEEE) is a faculty member at the Viterbi Department of Electrical and Computer Engineering, Technion–Israel Institute of Technology. His research interests lie at the intersection of the theoretical information sciences and the engineering of practical computing and storage systems. He has served on the technical program committees of leading conferences in both theory and systems. From 2010 to 2011, he was a Scientist at EPFL, the Swiss Federal Institute of Technology in Lausanne. From 2008 to 2010, he was a Research Staff Member at Hitachi Global Storage Technologies, San Jose Research Center. From 2018 to 2019, he held a Visiting Professor position at Western Digital Research, and a Visiting Scholar position at UC Berkeley. He received the B.Sc. degree in Electrical Engineering, summa cum laude, from the Technion in 2001, and the M.S. and Ph.D. degrees in Electrical Engineering from the California Institute of Technology, in 2004 and 2008, respectively. From 2000 to 2002, he was with Qualcomm, Israel R&D Center, where he worked on modeling, design and analysis in wireless communications. Dr. Cassuto has won the Best Student Paper Award in data storage from the IEEE Communications Society in 2010 as a student, and in 2019 as an adviser. He also won faculty awards from Qualcomm, Intel, and Western Digital. Before that, he won the 2001 Texas Instruments DSP and Analog Challenge $100,000 prize.

Yongjune Kim (Member, IEEE) received the B.S. (Hons.) and M.S. degrees in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2002 and 2004, respectively, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2016. From 2016 to 2018, he was a Post-doctoral Research Associate in the Coordinated Science Laboratory at University of Illinois at Urbana-Champaign, Urbana, IL, USA. From 2018 to 2020, he was with Western Digital Research, Milpitas, CA, USA. From 2020 to 2022, he was an Assistant Professor in the Department of Electrical Engineering and Computer Science, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea. Since 2022, he has been with the Department of Electrical and Engineering, Pohang University of Science and Technology (POSTECH), Pohang, South Korea, where he is currently an Assistant Professor. He is also affiliated with the Institute for Convergence Research and Education in Advanced Technology, Yonsei University, Seoul, South Korea. His research interests include semantic communications, information theory, and machine learning. He received the IEEE Data Storage Best Student Paper Award, the Best Paper Award of the 2016 IEEE International Conference on Communications (ICC), the Best Paper Award (honorable mention) of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), and the Best Paper Award of the 31st Samsung Semiconductor Technology Symposium.