Generalized Longest Repeated Substring Min-Entropy Estimator

Jiheon Woo*, Chanhee Yoo*, Young-Sik Kim†, and Yuval Cassuto‡, Yongjune Kim*
*Department of Electrical Engineering and Computer Science, DGIST, Daegu, South Korea
Email: {jhwoo1997, yoo1209, yjk}@dgist.ac.kr
†Department of Information and Communication Engineering, Chosun University, Gwangju, South Korea
Email: iamyskim@chosun.ac.kr
‡Viterbi Department of Electrical and Computer Engineering, Technion – Israel Institute of Technology, Haifa, Israel
Email: ycassuto@ee.technion.ac.il

***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 current-standard 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]. The *min-entropy* is a widely used randomness metric in cryptographic applications since it measures the difficulty of guessing the most likely output [4]–[6]. It is well known that the min-entropy is a lower bound on the Shannon entropy and the Rényi entropy, i.e., one of the most conservative entropy.

For independent and identically distributed (IID) sources, the min-entropy can be readily estimated by the empirical estimator [4]. However, it is difficult to estimate the min-entropy of non-IID sources accurately. Hence, the US National Institute of Standards and Technology (NIST) recommendation document SP 800-90B lists ten different min-entropy estimators for non-IID sources and then conservatively selects the minimum among these ten different values as the final estimate of the min-entropy.

Among these ten min-entropy estimators, the *longest repeated substring (LRS) estimator* is especially motivated toward finding non-randomness in long sequences, which is missed by other estimators. However, the metric used by the LRS estimator is the *collision entropy* (the Rényi entropy of order two) [4], and not the min-entropy. Since the collision entropy is only an upper bound on the min-entropy, the LRS estimator overestimates the min-entropy, which violates the conservative estimation methodology of NIST SP 800-90B.

In this paper, we propose two techniques to amend the LRS estimator for accurate min-entropy estimation. The first resolves the overestimation problem by providing an estimation of the min-entropy instead of the collision entropy. The second estimates the min-entropy using empirical statistics of Rényi entropies of general α order, instead of just $\alpha = 2$ as in the original estimator. In our main theoretical result, we show that higher orders can reduce the variance of the min-entropy estimates.

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.

The rest of this paper is organized as follows. Section II briefly reviews the LRS estimator of NIST SP 800-90B. Section III presents our modified LRS estimator that accurately estimates the min-entropy. Section IV presents the generalized LRS estimator and its analysis 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

Definition 1 (Min-Entropy): Suppose that the input sequence $\mathbf{s} = (s_1, \ldots, s_L)$ where $s_i \in \{x_1, \ldots, x_k\}$ is generated by a

given source S. Let $\mathbf{p} = (p_1, \ldots, p_k)$ denote the distribution of S. The min-entropy is defined as

$$H_\infty(S) = H_\infty(\mathbf{p}) = -\log_2 \theta, \tag{1}$$

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

Definition 2 (Power Sum): The power sum of order α (i.e., the α-th moment) for a distribution $\mathbf{p}$ is defined as $M_\alpha(\mathbf{p}) = \sum_{i=1}^{k} p_i^\alpha$.

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

Remark 4 (Collision Entropy): The power sum of order $\alpha = 2$, i.e., $M_2(\mathbf{p})$, is equivalent to the collision probability. The *collision* entropy is $H_2(\mathbf{p}) = -\log_2 M_2(\mathbf{p})$. It is well known that $H_2(\mathbf{p}) \geq H_\infty(\mathbf{p})$.

B. LRS Estimator and Its Overestimation Problem

For non-IID sources, NIST SP 800-90B proposes ten different min-entropy estimators [4]. These estimators independently perform their own estimations based on different statistics calculated from the examined non-IID sources. Among these ten estimators, the LRS estimator estimates the collision entropy based on the frequency of substrings (tuples) in the input sequence $\mathbf{s}$.

Algorithm 1 describes the LRS estimator of 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 $\{u, u+1, \ldots, v\}$.

Step 4 calculates the empirical collision probability of length-w substrings. Note that (2) is an *unbiased* estimator of the collision probability [7]. Step 5 computes the collision probability per sample (to normalize the entropies estimated from different lengths), 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.

The LRS estimator *overestimates* the min-entropy since it estimates the collision entropy instead of the min-entropy. Fig. 1(a) shows that the bias between the actual min-entropy and the estimate by the LRS estimator is considerable except for $p = 0.5$. Since NIST SP 800-90B conservatively selects the minimum among estimated values by ten estimators, the LRS estimator rarely contributes to the final estimate.

Algorithm 1 LRS estimator of NIST 800-90B [4]

Input: Sequence $\mathbf{s} = (s_1, \ldots, s_L)$ 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 $\mathbf{s}$ is less than 35.
2: Find the largest v such that the number of occurrences of the most common v-tuple in $\mathbf{s}$ 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:

$$P_w := \frac{\sum_i \binom{C_i}{2}}{\binom{l}{2}}, \tag{2}$$

where C_i is the number of occurrences of the ith unique w-tuple and l is the total number of w-tuples.
5: Compute the collision probability per sample:

$$\widetilde{P}_w := P_w^{1/w}. \tag{3}$$

6: **end for**
7: $\widehat{p}_c := \max\left\{\widetilde{P}_u, \ldots, \widetilde{P}_v\right\}$.
8: $\widetilde{p}_c := \min\left\{1, \widehat{p}_c + 2.576\sqrt{\frac{\widehat{p}_c(1-\widehat{p}_c)}{L-1}}\right\}$.
9: $H_2(S) := -\log_2 \widetilde{p}_c$.

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 aims to estimate the min-entropy instead of the collision entropy by using the collected statistics of the LRS estimator and the following bound.

Lemma 5 ([8, Theorem 6]): Suppose that $\theta = \max_{i\in\{1,\ldots,k\}}\{p_i\}$. Then, the following inequality holds:

$$H_\alpha(S) \leq \frac{1}{1-\alpha}\log_2\left(\theta^\alpha + \frac{(1-\theta)^\alpha}{(k-1)^{\alpha-1}}\right) \tag{4}$$

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

$$p_i = \begin{cases} \theta, & \text{if } i = 1; \\ \frac{1-\theta}{k-1}, & \text{otherwise.} \end{cases} \tag{5}$$

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

The bound (4) is the counterpart of Fano's inequality, which applies to the Shannon entropy.

Theorem 6: For the estimated collision probability $\widehat{p}_c$ by Algorithm 1, the following inequality holds:

$$\theta \leq \frac{\sqrt{(k-1)(p_c k - 1)} + 1}{k}, \tag{6}$$

where $p_c = \mathbb{E}(\widehat{p}_c)$. Since the near-uniform distribution achieves (4) with equality, (6) is the *sharp*[1] upper bound.

Proof: For $\alpha > 1$, (4) leads to

$$M_\alpha(\mathbf{p}) \geq \theta^\alpha + \frac{(1-\theta)^\alpha}{(k-1)^{\alpha-1}}. \tag{7}$$

Since $M_2(\mathbf{p})$ is equivalent to the collision probability p_c [7], we can set $p_c \geq \theta^2 + \frac{(1-\theta)^2}{(k-1)}$. Since $\theta = \max_{i\in\{1,\ldots,k\}}\{p_i\}$, it is clear that $\theta \geq \frac{1}{k}$. Since $\theta^2 + \frac{(1-\theta)^2}{(k-1)}$ is a non-decreasing function of θ for $\theta \geq \frac{1}{k}$, (6) holds. ■

Based on Theorem 6, we estimate $\widehat{\theta}$ as follows: $\widehat{\theta} = \frac{\sqrt{(k-1)(\widehat{p}_c k - 1)} + 1}{k}$, which is a conservative min-entropy estimation because an upper bound on θ leads to a lower bound on $H_\infty(S)$.

[1]The term "sharp bound" means that there exists a distribution that achieves this bound with equality.

Algorithm 2 Proposed LRS Estimator for the Min-Entropy

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

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

Algorithm 2 describes the proposed min-entropy LRS estimator. Step 1 of Algorithm 2 estimates the collision probability by using Algorithm 1. Theoretically, $p_c \geq \frac{1}{k}$ where the equality is achieved by the uniform distribution. If $\widehat{p}_c < \frac{1}{k}$, then we know that it results from estimation errors. Hence, in this case we set $\widehat{p}_c = \frac{1}{k}$, which leads to $\widehat{\theta} = \frac{1}{k}$. Step 7 ensures the confidence level of 99% as in Step 8 of Algorithm 1.

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. 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 distribution is 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, $\widehat{p}_c$ and $\widehat{\theta}$ instead of $\widetilde{p}_c$ and $\widetilde{\theta}$ are considered in our analysis.

We characterize the bias $b_{\text{proposed}}(S)$ by the *sharp* lower and upper bounds on θ for a given collision probability p_c. The sharp upper bound on θ is given in Theorem 6. We derive the sharp lower bound on θ by using the inverted near-uniform distribution. In [1], the inverted near-uniform distribution is defined as $\mathbf{p}_{\text{INU}}(\psi) = (p_1, \ldots, p_k)$ where

$$p_i = \begin{cases} \psi, & \text{if } i \in \left\{1, \ldots, \left\lfloor \frac{1}{\psi} \right\rfloor\right\}; \\ 1 - \left\lfloor \frac{1}{\psi} \right\rfloor \psi, & \text{if } i = \left\lfloor \frac{1}{\psi} \right\rfloor + 1; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Note that $\psi = \max\{\mathbf{p}_{\text{INU}}(\psi)\}$.

Lemma 7: For $\frac{1}{n+1} < \psi \leq \frac{1}{n}$ where $n \in \mathbb{N}$, the following relation holds: $\left\lfloor \frac{1}{\psi} \right\rfloor = \left\lfloor \frac{1}{M_2(\mathbf{p}_{\text{INU}}(\psi))} \right\rfloor = n$.

Proof: All proofs of Lemma, Theorem, and Corollary are given in this paper's longer version [9]. ∎

Algorithm 3 Generalized LRS estimator

Input: Sequence $\mathbf{s} = (s_1, ..., s_L)$ and an integer $\alpha \geq 2$
Output: Min-entropy $H_\infty(S)$.

1: Find the smallest u such that the number of occurrences of the most common u-tuple in $\mathbf{s}$ is less than 35.
2: Find the largest v such that the number of occurrences of the most common v-tuple in $\mathbf{s}$ is at least α.
3: **for** $w \in \{u, u+1, \ldots, v\}$ **do**
4: $\quad$ Estimate the w-tuple power sum of order α:

$$\widehat{M}_{\alpha,w} := \frac{\sum_i \binom{C_i}{\alpha}}{\binom{l}{\alpha}}, \tag{12}$$

$\quad$ where C_i is the number of occurrences of the ith unique w-tuple and l is the total number of w-tuples.
5: $\quad \widetilde{M}_{\alpha,w} := \widehat{M}_{\alpha,w}^{\frac{1}{w}}$.
6: **end for**
7: $\widetilde{M}_\alpha := \max\{\widetilde{M}_{\alpha,u}, \ldots, \widetilde{M}_{\alpha,v}\}$.
8: **if** $\widetilde{M}_\alpha > \frac{1}{k^{\alpha-1}}$ **then**
9: $\quad$ By the bisection method, solve the following equation for $\widehat{\theta} \in \left[\frac{1}{k}, 1\right]$:

$$\widetilde{M}_\alpha = \widehat{\theta}^\alpha + \frac{(1-\widehat{\theta})^\alpha}{(k-1)^{\alpha-1}}. \tag{13}$$

10: **else**
11: $\quad \widehat{\theta} := \frac{1}{k}$.
12: **end if**
13: $\widetilde{\theta} := \min\left(1, \widehat{\theta} + 2.576\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{L-1}}\right)$.
14: $H_\infty(S) := -\log_2 \widetilde{\theta}$.

Theorem 8: For any distribution $\mathbf{p} = (p_1, \ldots, p_k)$ with $n = \left\lfloor \frac{1}{p_c} \right\rfloor$, the following inequalities hold: $\psi \leq \theta \leq \breve{\theta}$, where

$$\psi = \frac{\sqrt{n\{p_c(n+1) - 1\}} + n}{n(n+1)}, \tag{9}$$

$$\breve{\theta} = \frac{\sqrt{(k-1)(p_c k - 1)} + 1}{k}. \tag{10}$$

For given p_c and k, we define the *estimation gap* of θ as

$$g(p_c, k) = \breve{\theta} - \psi, \tag{11}$$

which is the maximum possible bias. The following theorem shows that the estimation gap increases with k.

Theorem 9: For non-deterministic sources, the estimation gap $g(p_c, k) = \breve{\theta} - \psi$ increases with k.

Corollary 10: For binary sources with $k = 2$, the estimation gap is zero, i.e., $g(p_c, k = 2) = 0$.

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 α increases beyond 2.

The generalized LRS estimator is described in Algorithm 3. First, it estimates the power sum $M_\alpha(\mathbf{p})$ for a given α by Steps 1–7. Step 2 of Algorithm 3 is modified from Algorithm 1 to estimate $M_\alpha(\mathbf{p})$. Step 4 estimates the w-tuple power sum of order α by counting the α-wise collisions. Step 5 computes the power sum of order α per sample (to normalize the estimated min-entropy) and Step 7 conservatively chooses the maximum among estimated power sums of α, which is denoted by $\widetilde{M}_\alpha$.

The key modification needed for allowing general α is the transformation in Step 9 from $\widetilde{M}_\alpha$ (i.e., the conservative estimate of the power sum) to $\widehat{\theta}$. In the special case of $\alpha = 2$ this transformation was given in the closed form (Step 3 in Algorithm 2), whereas for general α we need to find $\widehat{\theta}$ via bisection. In the special case $\widetilde{M}_\alpha \leq \frac{1}{k^{\alpha-1}}$ (branching to Line 10), the power-sum estimate is equal or lower than the minimum value attained by the uniform distribution, hence we set $\widehat{\theta}$ to the minimum possible value of $\frac{1}{k}$.

Similar to Algorithm 2, Algorithm 3 solves the bias problem of Algorithm 1. We show next that Algorithm 3 also offers an advantage over Algorithm 2 in reducing the estimation *variance*. 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 following theorem analyzes the estimation variance (for uniform sources) and shows that it is decreasing with α. Empirical evidence of this behavior is shown in Section V for more general sources.

Theorem 11: For a uniformly distributed $\mathbf{s} = (s_1, \ldots, s_L)$ with a large L, the variance ratio's dependence on α is as follows:

$$\xi(\alpha) = \frac{\mathsf{Var}(\widehat{\theta}_{\alpha+1})}{\mathsf{Var}(\widehat{\theta}_\alpha)} \approx \left(\frac{\alpha}{\alpha+1}\right)^4, \tag{14}$$

where $\widehat{\theta}_\alpha$ and $\widehat{\theta}_{\alpha+1}$ are the estimated $\widehat{\theta}$ in Algorithm 3 by using $\widetilde{M}_\alpha$ and $\widetilde{M}_{\alpha+1}$, respectively, and $\approx$ hides multiplicative terms that tend to 1 as L goes to infinity.

Proof: We denote the number of α-wise collisions as $D_{\alpha,w}$ for the w-tuples in Step 4 of Algorithm 3, which is given by $D_{\alpha,w} = \sum_{i=1}^{k^w} \binom{C_i}{\alpha}$, where C_i is the number of occurrences of the i-th w-tuple. We suppose that $\binom{C_i}{\alpha} = 0$ if $C_i < \alpha$. For every subset $I \subseteq \{1, \ldots, l = \lfloor \frac{L}{w} \rfloor\}$ of size α, we define X_I to be a 0-1 random variable that gets the value 1 iff all the values x_i are the same (i.e., I forms a α-wise collision). It is clear that $D_{\alpha,w} = \sum_{|I|=\alpha} X_I$ and $\mathbb{E}(X_I) = M_{\alpha,w}$, where $M_{\alpha,w}$ is the w-tuple power sum of order α. Also, we set $\overline{X}_I = X_I - M_{\alpha,w}$ as in [10].

For two subsets I and J such that $|I| = |J| = \alpha$, $\mathbb{E}(\overline{X}_I \cdot \overline{X}_J) = \mathbb{E}(\overline{X}_I) \cdot \mathbb{E}(\overline{X}_J) = 0$ 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, $\mathbb{E}(\overline{X}_I \cdot \overline{X}_J) = M_{\alpha+t,w} - M_{\alpha,w}^2$ if $|I \cup J| = \alpha + t < 2\alpha$ [10]. Since $M_{\alpha,w} = \frac{1}{k^{w(\alpha-1)}}$ for a uniformly distributed source, we obtain $\mathbb{E}(\overline{X}_I \cdot \overline{X}_J) = \frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}}$.

The variance of $D_{\alpha,w}$ is given by

$$\mathsf{Var}(D_{\alpha,w}) = \sum_{t=0}^{\alpha-1} \sum_{|I\cup J|=\alpha+t} \mathbb{E}(\overline{X}_I \cdot \overline{X}_J) \tag{15}$$

$$\approx \frac{1}{k^{w(\alpha-1)}} \binom{l}{\alpha} \sum_{t=0}^{\alpha-2} \binom{l}{t} \binom{\alpha}{t} \left(\frac{1}{k^{wt}} - \frac{1}{k^{w(\alpha-1)}}\right). \tag{16}$$

By taking into account normalization in Step 5 of Algorithm 3, we obtain

$$\mathsf{Var}(\widetilde{M}_{\alpha,w}) \approx \frac{1}{w^2} \cdot \mathbb{E}(\widehat{M}_{\alpha,w})^{\frac{2(1-w)}{w}} \cdot \mathsf{Var}(\widehat{M}_{\alpha,w}) \tag{17}$$

$$= \frac{1}{w^2} \cdot k^{2(\alpha-1)(w-1)} \cdot \frac{\mathsf{Var}(D_{\alpha,w})}{\binom{l}{\alpha}^2} \tag{18}$$

$$\approx \frac{k^{(\alpha-1)(w-2)}}{w^2} \cdot \frac{\sum_{t=0}^{\alpha-2} \binom{l}{t}\binom{\alpha}{t}\left(k^{-wt} - k^{-w(\alpha-1)}\right)}{\binom{l}{\alpha}}. \tag{19}$$

In Step 7 of Algorithm 3, the maximum among $\{\widetilde{M}_{\alpha,u}, \ldots, \widetilde{M}_{\alpha,v}\}$ is chosen as $\widetilde{M}_\alpha$. It is difficult to characterize which $\widetilde{M}_{\alpha,w}$ for $w \in \{u, \ldots, v\}$ is the maximum value. As a conservative approach, we set $\mathsf{Var}(\widetilde{M}_\alpha) \approx \mathsf{Var}(\widetilde{M}_{\alpha,\overline{v}})$ where $\overline{v}$ 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 α times in the sequence (see Step 2 in Algorithm 3). Then,

$$\mathsf{Var}(\widetilde{M}_\alpha) \approx \frac{k^{(\alpha-1)(\overline{v}_\alpha-2)}}{\overline{v}_\alpha^2} \cdot \frac{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t}\binom{\alpha}{t}\left(k^{-t\overline{v}_\alpha} - k^{-(\alpha-1)\overline{v}_\alpha}\right)}{\binom{l_\alpha}{\alpha}}, \tag{20}$$

where we denote $\overline{v} = \overline{v}_\alpha$ and $l = l_\alpha$ since they depend on α. Note that $\overline{v}_\alpha \approx \frac{1}{\alpha-1}\log_k \binom{l_\alpha}{\alpha}$ is derived from the cutoff condition $\mathbb{E}(D_{\alpha,1}) \cdot (M_\alpha)^{\overline{v}-1} \geq 1$ and $\mathbb{E}(D_{\alpha,1}) \cdot (M_\alpha)^{\overline{v}} < 1$ where the left-hand sides of the inequalities equal (proof omitted) $\mathbb{E}(D_{\alpha,\overline{v}})$ and $\mathbb{E}(D_{\alpha,\overline{v}+1})$, respectively, that are, the expected numbers of α-repeating tuples of the corresponding lengths.

By Taylor approximation, $\mathsf{Var}(\widehat{\theta}_\alpha) \approx z(\widehat{\theta}_\alpha, \alpha)^2 \cdot \mathsf{Var}(\widetilde{M}_\alpha)$ where $z(\widehat{\theta}_\alpha, \alpha) = \frac{d\widehat{\theta}_\alpha}{d\widetilde{M}_\alpha} = \frac{1}{\alpha\left\{\widehat{\theta}_\alpha^{\alpha-1} - \left(\frac{1-\widehat{\theta}_\alpha}{k-1}\right)^{\alpha-1}\right\}}$, which is derived from (13). Then, we obtain $\frac{z(\widehat{\theta}_{\alpha+1},\alpha+1)}{z(\widehat{\theta}_\alpha,\alpha)} \approx \frac{\alpha-1}{\alpha+1} \cdot k$. Finally, skipping some technical steps,

$$\xi(\alpha) = \frac{\mathsf{Var}(\widehat{\theta}_{\alpha+1})}{\mathsf{Var}(\widehat{\theta}_\alpha)} \approx \frac{z(\widehat{\theta}_{\alpha+1},\alpha+1)^2}{z(\widehat{\theta}_\alpha,\alpha)^2} \cdot \frac{\mathsf{Var}(\widetilde{M}_{\alpha+1})}{\mathsf{Var}(\widetilde{M}_\alpha)} \tag{21}$$

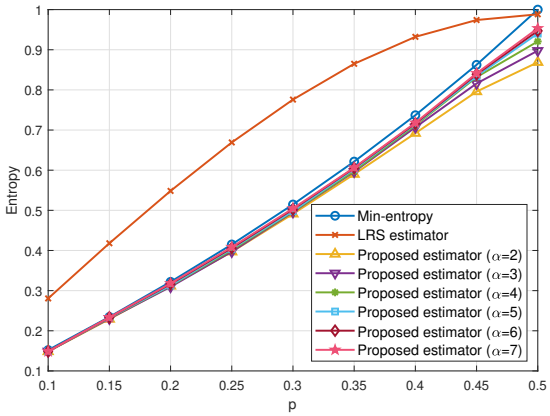
$$\approx \left(\frac{\alpha}{\alpha+1}\right)^4 \cdot \frac{\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t}\binom{\alpha+1}{t}\binom{l_{\alpha+1}}{\alpha+1}^{-\frac{t}{\alpha}}}{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t}\binom{\alpha}{t}\binom{l_\alpha}{\alpha}^{-\frac{t}{\alpha-1}}}. \tag{22}$$

For a large L, (22) converges to $\left(\frac{\alpha}{\alpha+1}\right)^4$. The detailed proof is given in [9]. ∎

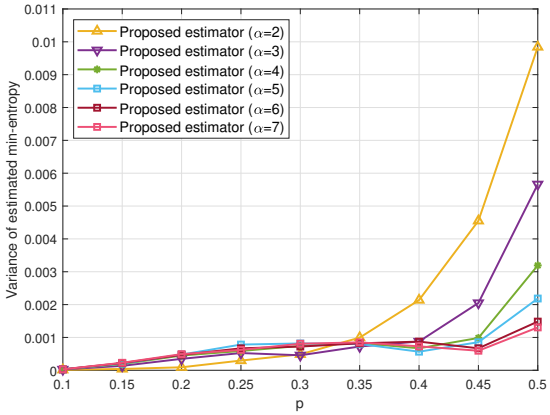
Since $\xi(\alpha) < 1$, $\mathsf{Var}(\widehat{\theta})$ decreases with α for high-entropy sources. Thus, the generalized LRS estimator can provide stable min-entropy estimates.

V. Numerical Results

We evaluate our proposed estimators for the following representative data samples. The more extensive numerical results are provided in [9].

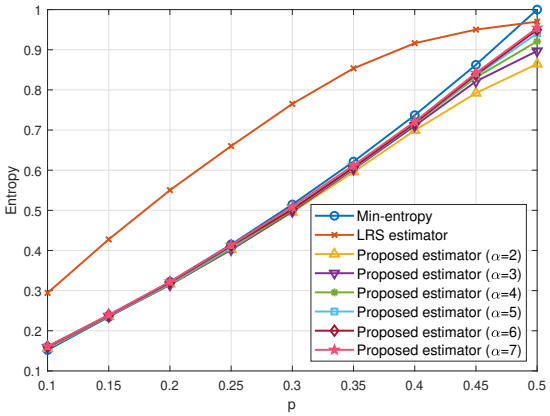


(a)

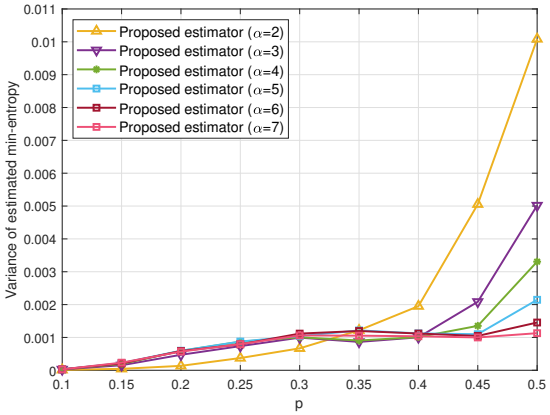


(b)

Fig. 1. (a) Estimated min-entropy and (b) the variance of min-entropy estimates by the proposed generalized LRS estimator for the BMS sources with p.



(a)



(b)

Fig. 2. (a) Estimated min-entropy and (b) the variance of min-entropy by the proposed generalized LRS estimator for the first-order Markov sources with $p = p(1|0) = p(0|1)$.

- *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 $P(S_{i+1} = 1|S_i = 0) = P(S_{i+1} = 0|S_i = 1) = p$ (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.

Fig. 1 compares the min-entropy estimators for BMS as a function of p. The original LRS estimator estimates the collision entropy instead of the min-entropy, and this is reflected in the significant overestimation it exhibits in Fig. 1(a). In the same plot, the proposed estimator shows much more accurate estimates. It can be seen that the larger-α estimators give more accurate min-entropy estimates, and that is thanks to their use of higher-order repeat statistics which better capture the infinite-order min-entropy. In Fig. 1(b), we observe that as $p \to 0.5$ (i.e., uniformly distributed sources), the higher α also reduces $\mathsf{Var}(\widehat{\theta})$, which supports Theorem 11. We note that the reduction of $\mathsf{Var}(\widehat{\theta})$ diminishes as α increases as shown in Fig. 1(b).

For the first-order Markov sources, the min-entropy estimators estimate the min-entropy rate. By [11], [12], 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. Fig. 2 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. 2(a). The proposed estimator effectively improves the accuracy of min-entropy estimates.

VI. Conclusion

We proposed accurate min-entropy estimators to resolve the overestimation problem of the LRS estimator. Although the first 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 estimator by parameterizing α instead of restricting to $\alpha = 2$. It is shown that the generalized LRS estimator can improve the bias and variance of min-entropy estimates.

REFERENCES

[1] P. Hagerty and T. Draper, "Entropy bounds and statistical tests," in *Proc. NIST Random Bit Generation Workshop*, Dec. 2012, pp. 1–28.

[2] W. Killmann and W. Schindler, *A proposal for: Functionality classes for random number generators*, German Federal Office for Information Security (BSI) Std., Rev. 2, Sep. 2011.

[3] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo, *A statistical test suite for random and pseudorandom number generators for cryptographic applications*, NIST Special Publication 800-22 Std., Rev. 1a, Apr. 2010.

[4] M. S. Turan, E. Barker, J. Kelsey, K. A. McKay, M. L. Baish, and M. Boyle, *Recommendation for the Entropy Sources Used for Random Bit Generation*, NIST Special Publication 800-90B Std., Jan. 2018.

[5] J. Kelsey, K. A. McKay, and M. S. Turan, "Predictive models for min-entropy estimation," in *Proc. Int. Workshop Cryptograph. Hardw. Embedded Syst. (CHES)*, Berlin, Heidelberg, Sep. 2015, pp. 373–392.

[6] Y. Kim, C. Guyot, and Y.-S. Kim, "On the efficient estimation of min-entropy," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3013–3025, Apr. 2021.

[7] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating Rényi entropy," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, Jan. 2015, pp. 1855–1869.

[8] M. Ben-Bassat and J. Raviv, "Renyi's entropy and the probability of error," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 324–331, May 1978.

[9] J. Woo, C. Yoo, Y.-S. Kim, Y. Cassuto, and Y. o. Kim, "Generalized LRS estimator for min-entropy estimation," *arXiv preprint arXiv:2112.09376*, 2021. [Online]. Available: http://arxiv.org/abs/2112.09376

[10] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, "Sampling algorithms: Lower bounds and applications," in *Proc. Annu. ACM Symp. Theory Comput. (STOC)*, Feb. 2002, pp. 266–275. [Online]. Available: https://webee.technion.ac.il/people/zivby/papers/sampling/sampling_full.ps

[11] Z. Rached, F. Alajaji, and L. Lorne Campbell, "Renyi's divergence and entropy rates for finite alphabet Markov sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1553–1561, May 2001.

[12] S. Kamath and S. Verdú, "Estimation of entropy rate and Rényi entropy rate for Markov chains," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 685–689.