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.

J. Woo and C. Yoo contributed equally. J. Woo, C. Yoo, and Y. Kim are with the Department of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, South Korea (e-mail: {jhwoo1997, yoo1209, yjk}@dgist.ac.kr). Y.-S. Kim is with the Department of Information and Communication Engineering, Chosun University, Gwangju 61452, South Korea (e-mail: iamyskim@chosun.ac.kr). Y. Cassuto is with the Viterbi Department of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: ycassuto@ee.technion.ac.il).

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.

Among the 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.

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 [7, 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 LRS estimator by parameterizing the order, i.e., α of the Rényi entropy. By adopting a higher order α 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 α, although the reduction of the variance diminishes as α 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

TABLE I
CLASSIFICATION OF NIST SP 800-90B ESTIMATORS [1], [8]

Statistic-based estimator [2]	Prediction-based estimator [3]
Most common value estimator	MultiMCW prediction estimator
Collision estimator	Lag prediction estimator
Markov estimator	MultiMMC prediction estimator
Compression estimator	LZ78Y prediction estimator
t-Tuple estimator	
LRS estimator	

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

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

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

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

$$H_\alpha(S) = H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{k} p_i^\alpha. \tag{2}$$

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

$$H_2(S) = H_2(\mathbf{p}) = -\log_2 \sum_{i=1}^{k} p_i^2. \tag{3}$$

The min-entropy is defined as

$$H_\infty(S) = H_\infty(\mathbf{p}) = -\log_2 \theta, \tag{4}$$

where

$$\theta = \max_{i \in \{1, \ldots, k\}} \{p_i\}. \tag{5}$$

Definition 1 (Power Sum): The power sum of order α (i.e., the αth moment) for a distribution $\mathbf{p}$ is defined as

$$M_\alpha(\mathbf{p}) = \sum_{i=1}^{k} p_i^\alpha. \tag{6}$$

Remark 2 (Collision Probability): The power sum of order $\alpha = 2$ (i.e., $M_2(\mathbf{p})$) is equivalent to the collision probability,

---

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

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:

$$P_w := \frac{\sum_i \binom{C_i}{2}}{\binom{l}{2}}, \tag{9}$$

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:

$$\widetilde{P}_w := P_w^{1/w}. \tag{10}$$

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: $H_2(S) := -\log_2 \widetilde{p}_c.$

---

which is the probability that two arbitrary source outputs are equal.

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

Remark 4: The following relations are well known:

$$H(S) = \lim_{\alpha \to 1} H_\alpha(S), \tag{7}$$

$$H_\infty(S) = \lim_{\alpha \to \infty} H_\alpha(S). \tag{8}$$

Remark 5: The Rényi entropy is non-increasing in α [9]. 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. 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 $\mathbf{s}$. 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], [8].

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 set 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 $l = \lfloor \frac{L}{w} \rfloor$.

The collision probability estimation by (9) is a key step of the LRS estimator, which was considered in [10], [11] for testing whether a distribution is close to uniform. Note that (9) is an *unbiased* estimator of the collision probability [12].

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. 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 from 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 [13, Table 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 would not affect the final estimate. Since the LRS estimator computes based on larger sizes of susbtrings than the t-tuple estimator, the LRS estimator takes around ten times longer execution time than the t-tuple estimator [8, Table V]. In spite of this considerable execution time, the LRS estimator rarely affects the final min-entropy estimate due to overestimation problem noted above.

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

$$H_\alpha(S) \leq \frac{1}{1 - \alpha} \log_2 \left(\theta^\alpha + \frac{(1 - \theta)^\alpha}{(k - 1)^{\alpha - 1}} \right) \quad (11)$$

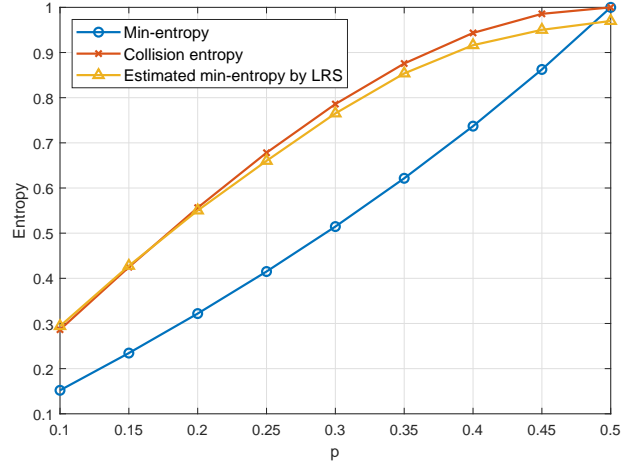

Fig. 1. Comparison of min-entropy, collision entropy and estimated value by LRS estimator for binary memoryless source (BMS) with parameter p. While the LRS estimator has a small underestimation gap to the true collision entropy (due to Step 8 in Algorithm 1), its overestimation gap (bias) to the min-entropy is significant.

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

$$p_i = \begin{cases} \theta, & \text{if } i = 1; \\ \frac{1 - \theta}{k - 1}, & \text{otherwise.} \end{cases} \quad (12)$$

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 $\widehat{p}_c$ by Algorithm 1, the following inequality holds:

$$\theta \leq \frac{\sqrt{(k - 1)(p_c k - 1)} + 1}{k}, \quad (13)$$

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

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

$$M_\alpha(\mathbf{p}) \geq \theta^\alpha + \frac{(1 - \theta)^\alpha}{(k - 1)^{\alpha - 1}}. \quad (14)$$

Since $M_2(\mathbf{p})$ is equivalent to the collision probability p_c [12], we can set $p_c \geq \theta^2 + \frac{(1 - \theta)^2}{(k - 1)}$. By (5), 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 $\theta \geq \frac{1}{k}$, (13) holds. ∎

Based on Theorem 7, we estimate $\widehat{\theta}$ as follows:

$$\widehat{\theta} = \frac{\sqrt{(k - 1)(\widehat{p}_c k - 1)} + 1}{k}, \quad (15)$$

which is a conservative min-entropy estimation. It is because an upper bound on θ leads to a lower bound on $H_\infty(S)$, i.e.,

$$H_\infty(S) = -\log_2 \theta \geq -\log_2 \widehat{\theta} \quad (16)$$

if $\widehat{p}_c = p_c$.

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 2 Proposed Estimator (Improved LRS Estimator)

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:     $\widehat{\theta} := \frac{\sqrt{(k-1)(\widehat{p}_c k-1)}+1}{k}$.
  4: **else**
  5:     $\widehat{\theta} := \frac{1}{k}$.
  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 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. 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, $\widehat{p}_c$ and $\widehat{\theta}$ instead of $\widetilde{p}_c$ and $\widetilde{\theta}$ are considered in our analysis.

The bias of the LRS estimator is given by

$$b_{\text{LRS}}(S) = \mathbb{E}(-\log_2 \widehat{p}_c) - H_\infty(S). \quad (17)$$

Proposition 8: The LRS estimator is *overestimating*, i.e., $b_{\text{LRS}}(S) > 0$.

Proof: We show that $b_{\text{LRS}}(S) > 0$ as follows:

$$\begin{aligned}
b_{\text{LRS}}(S) &= \mathbb{E}(-\log_2 \widehat{p}_c) - H_\infty(S) \\
&> -\log_2 \mathbb{E}(\widehat{p}_c) - H_\infty(S) \quad (18) \\
&= H_2(S) - H_\infty(S) \quad (19) \\
&\geq 0 \quad (20)
\end{aligned}$$

where (18) follows from Jensen's inequality. Since $-\log x$ is strictly convex and $\widehat{p}_c$ is not constant for non-deterministic sources (18) holds. Also, (20) follows from Remark 5. Hence, $b_{\text{LRS}}(S) > 0$. ∎

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

$$b_{\text{proposed}}(S) = \mathbb{E}(-\log_2 \widehat{\theta}) - H_\infty(S) > \log_2 \frac{\theta}{\mathbb{E}(\widehat{\theta})}, \quad (21)$$

where $\log_2 \frac{\theta}{\mathbb{E}(\widehat{\theta})} \leq 0$ since $\widehat{\theta}$ is an estimate of the upper bound on θ (see Theorem 7). If $\mathbb{E}(-\log_2 \widehat{\theta}) \simeq -\log_2 \mathbb{E}(\widehat{\theta})$, 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 $b_{\text{proposed}}(S)$ by the *sharp*[1] lower and upper bounds on θ for a given collision probability p_c. The sharp upper bound on θ is given in Theorem 7. We derive the sharp lower bound on θ by using the inverted near-uniform distribution. In [2], the inverted near-uniform distribution is defined as $\mathbf{p}_{\text{INU}}(\psi) = (p_1, \ldots, p_k)$ where

$$p_i = \begin{cases} \psi, & \text{if } i \in \left\{1, \ldots, \left\lfloor \frac{1}{\psi} \right\rfloor\right\}; \\ 1 - \left\lfloor \frac{1}{\psi} \right\rfloor \psi, & \text{if } i = \left\lfloor \frac{1}{\psi} \right\rfloor + 1; \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

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

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

$$\left\lfloor \frac{1}{\psi} \right\rfloor = \left\lfloor \frac{1}{M_2(\mathbf{p}_{\text{INU}}(\psi))} \right\rfloor = n. \quad (23)$$

Proof: If $\psi = \frac{1}{n}$, then

$$M_2(\mathbf{p}_{\text{INU}}(\psi)) = \frac{1}{n}. \quad (24)$$

Hence, (23) holds.

If $\frac{1}{n+1} < \psi < \frac{1}{n}$, then $M_2(\mathbf{p}_{\text{INU}}(\psi))$ is an increasing function of ψ. It is because $\frac{dM_2(\mathbf{p}_{\text{INU}}(\psi))}{d\psi} = 2n(n+1)\{\psi - \frac{1}{n+1}\} > 0$. By (24), we obtain $\frac{1}{n+1} < M_2(\mathbf{p}_{\text{INU}}) < \frac{1}{n}$. 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 $\psi \simeq M_2(\mathbf{p}_{\text{INU}})$. If $\psi = \frac{1}{n}$, then the collision entropy is the same as the min-entropy.

Theorem 11: For any distribution $\mathbf{p} = (p_1, \ldots, p_k)$ with $n = \left\lfloor \frac{1}{p_c} \right\rfloor$, the following inequalities hold:

$$\psi \leq \theta \leq \widehat{\theta} \quad (25)$$

where

$$\psi = \frac{\sqrt{n\{p_c(n+1)-1\}}+n}{n(n+1)}, \quad (26)$$

$$\widehat{\theta} = \frac{\sqrt{(k-1)(p_c k-1)}+1}{k}. \quad (27)$$

---

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

Proof: Since $\widehat{\theta}$ is derived in Theorem 7, we need to derive only the sharp lower bound ψ. The lower bound ψ is achieved with equality by the inverted near-uniform distribution $\mathbf{p}_{\mathrm{INU}}(\psi)$ [2], [14]. Hence, we need to identify $\mathbf{p}_{\mathrm{INU}}(\psi)$ satisfying $M_2(\mathbf{p}_{\mathrm{INU}}(\psi)) = p_c$. Suppose that $\frac{1}{n+1} < \psi \leq \frac{1}{n}$ where $n \in \mathbb{N}$ (i.e., $\left\lfloor \frac{1}{\psi} \right\rfloor = n$). By (22) and Lemma 9, we obtain $M_2(\mathbf{p}_{\mathrm{INU}}(\psi)) = n\psi^2 + (1 - n\psi)^2 = p_c$, 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

$$-\log_2 \widehat{\theta} \leq H_\infty(\mathbf{p}) \leq -\log_2 \psi, \tag{28}$$

where $H_\infty(\mathbf{p}) = -\log_2 \theta$.

We note that ψ depends only on p_c because $n = \left\lfloor \frac{1}{p_c} \right\rfloor$. On the other hand, $\widehat{\theta}$ depends on p_c and k (alphabet size $|S|$). For given p_c and k, we define the *estimation gap* of θ as

$$g(p_c, k) = \widehat{\theta} - \psi, \tag{29}$$

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 $g(p_c, k) = \widehat{\theta} - \psi$ increases with k.

Proof: Since ψ does not depend on k, we show that $\widehat{\theta}(k)$ is an increasing function of k. The derivative of $\widehat{\theta}(k)$ is given by

$$\frac{d\widehat{\theta}(k)}{dk} = \frac{p_c k + k - 2 - 2\sqrt{(k-1)(p_c k - 1)}}{2k^2\sqrt{(k-1)(p_c k - 1)}}. \tag{30}$$

We can set $p_c > \frac{1}{k}$ because $p_c = \frac{1}{k}$ means that $\mathbf{p}$ is the uniform distribution, i.e., $\psi = \widehat{\theta} = \frac{1}{k}$ and $H_\infty(\mathbf{p}) = \log_2 k$. By the arithmetic-geometric mean inequality,

$$p_c k + k - 2 = (p_c k - 1) + (k - 1) \geq 2\sqrt{(k-1)(p_c k - 1)}. \tag{31}$$

Hence, $\frac{d\widehat{\theta}(k)}{dk} > 0$ 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 $\theta = \frac{1}{2}$, $\mathbf{p}$ 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., $\theta = \psi = \widehat{\theta} = \frac{\sqrt{2p_c - 1} + 1}{2}$. ∎

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 improves the accuracy 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 θ, respectively. Also, it shows that $\psi = \widehat{\theta}$ for $k = 2$, i.e., the proposed estimator is unbiased for binary souces. 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.

We note that the final min-entropy estimation could not be perfectly unbiased because of the two steps of the original LRS
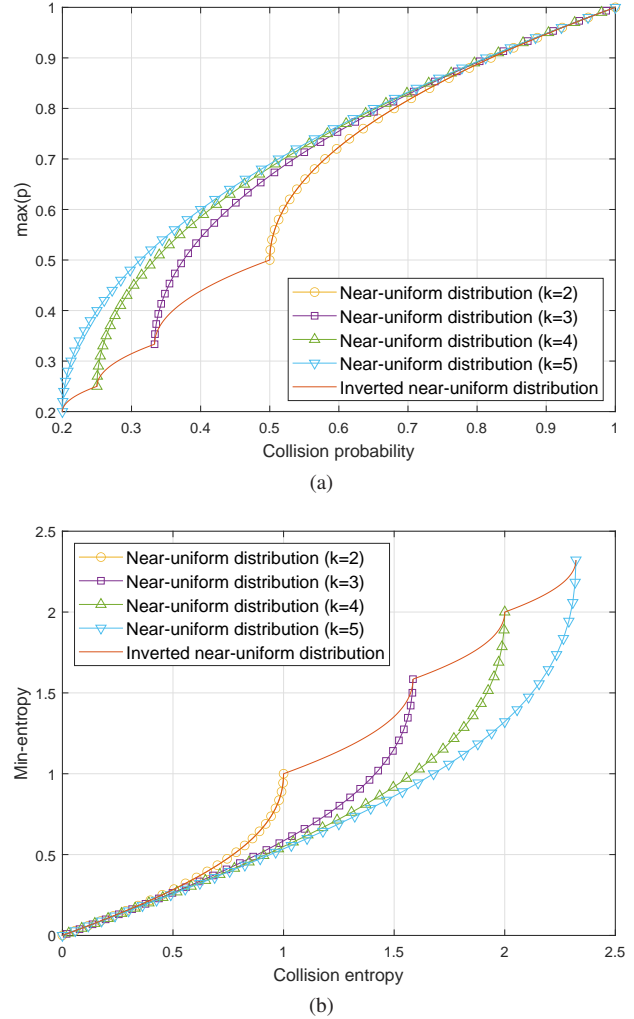

(a)

(b)

Fig. 2. The relation of (a) the collision probability p_c and $\theta = \max_{i \in \{1,\dots,k\}} p_i$ and (b) the collision entropy and the min-entropy for $\mathbf{p}_{\mathrm{NU}}$ and $\mathbf{p}_{\mathrm{INU}}$.

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 α increases beyond 2.

A. Generalized LRS Estimator

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

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:     Estimate the w-tuple power sum of order α:

$$\widehat{M}_{\alpha,w} := \frac{\sum_i \binom{C_i}{\alpha}}{\binom{l}{\alpha}}, \tag{32}$$

   where C_i is the number of occurrences of the ith unique w-tuple and l is the total number of w-tuples.
5:     $\widetilde{M}_{\alpha,w} := \widehat{M}_{\alpha,w}^{\frac{1}{w}}$.
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:     By the bisection method, solve the following equation for $\widehat{\theta} \in \left[\frac{1}{k}, 1\right]$:

$$\widetilde{M}_\alpha = \widehat{\theta}^\alpha + \frac{(1-\widehat{\theta})^\alpha}{(k-1)^{\alpha-1}}. \tag{33}$$

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

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 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 estimation $\widehat{M}_\alpha(\mathbf{p})$ by (32) is a key step, which generalizes (9) in Algorithm 1. The estimation by (32) is unbiased [12], [15]. However, $\widetilde{M}_\alpha$ in Step 7 is an overestimate of $M_\alpha(\mathbf{p})$, 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 $\widehat{\theta}$ from $\widetilde{M}_\alpha$ in Steps 8–12. First, we note that $\widetilde{M}_\alpha \geq \frac{1}{k^{\alpha-1}}$ where the equality is achieved by the uniform distribution. Hence, we set $\widehat{\theta} = \frac{1}{k}$ in Step 11 if $\widetilde{M}_\alpha < \frac{1}{k^{\alpha-1}}$. If $\widetilde{M}_\alpha > \frac{1}{k^{\alpha-1}}$, then $\widehat{\theta}$ is estimated by (14), which is the sharp upper bound on θ. Especially, $\widehat{\theta}$ is unbiased for binary sources $k = 2$ (see Corollary 14). Step 13 ensures the confidence level of 99 % under the Gaussian assumption. Finally, Step 14 estimates the min-entropy from $\widetilde{\theta}$.

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 α affects the variance of estimation.

B. Variance of Generalized LRS Estimator

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

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

Theorem 16: For a uniformly distributed $\mathbf{s} = (s_1, \ldots, s_L)$ with a large L, the variance ratio's dependence on α is as follows:

$$\xi(\alpha) = \frac{\text{Var}(\widehat{\theta}_{\alpha+1})}{\text{Var}(\widehat{\theta}_\alpha)} \approx \left(\frac{\alpha}{\alpha+1}\right)^4, \tag{34}$$

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

Proof: The proof is given in Appendix B. ■
Since $\xi(\alpha) < 1$, $\text{Var}(\widehat{\theta})$ decreases with α for high-entropy sources. The reduction of $\text{Var}(\widehat{\theta}_\alpha)$ diminishes as α increases.

The range of $w \in \{u, \ldots, v\}$ is an important parameter that affects the variance of $\widehat{\theta}$. It is clear that

$$v_\alpha \geq v_{\alpha+1}, \quad u = u_\alpha = u_{\alpha+1}, \tag{35}$$

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

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

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

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], [16]:

- *Binary memoryless source (BMS):* Samples are generated by Bernoulli distribution with $P(S = 1) = p$ and $P(S = 0) = 1 - p$ (IID);
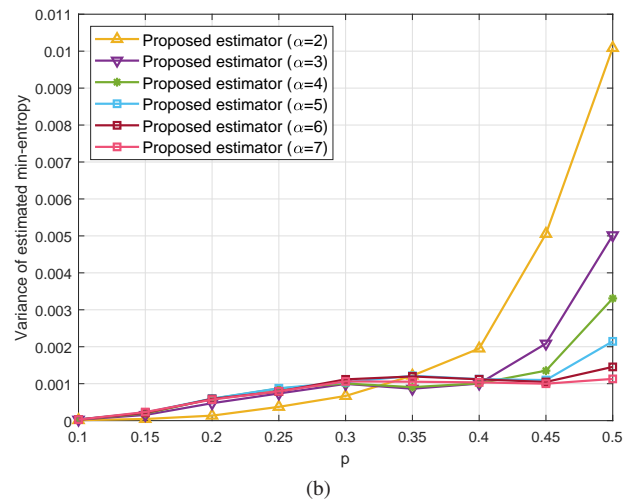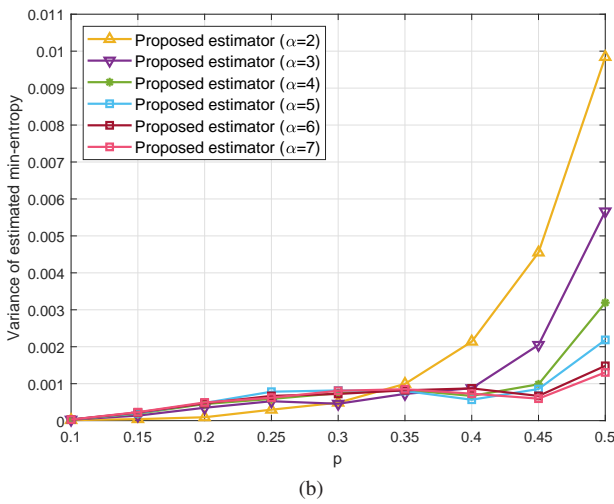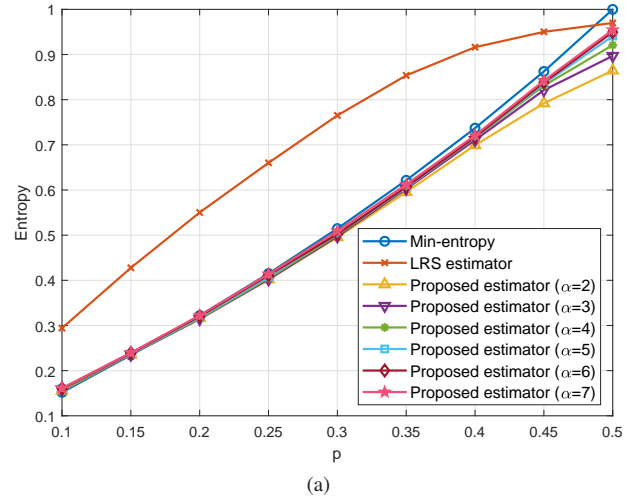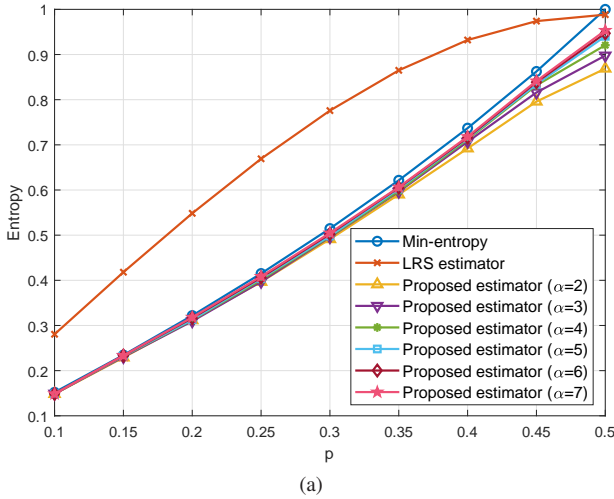
Fig. 3. (a) Estimated min-entropy and (b) the variance of min-entropy estimates by the proposed generalized LRS estimator for the BMS sources with p.

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 $p = p(1|0) = p(0|1)$.

- *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);
- *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).

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 $H_2(S) = -\log_2\{p^2 + (1-p)^2\}$, 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 α reduces $\text{Var}(\widehat{\theta})$, which supports Theorem 16. We note that the reduction of $\text{Var}(\widehat{\theta})$ diminishes as α increases as shown in Fig. 3(b).

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

It is worth mentioning that because of the finite sample size L, the higher order α reduces the number of valid C_i due to the requirement of $C_i \geq \alpha$ in (32). Then, (32) could underestimate the power sum of α. Given the sample size L, an α value should be picked so as to satisfy $C_i \geq \alpha$ for values of w as
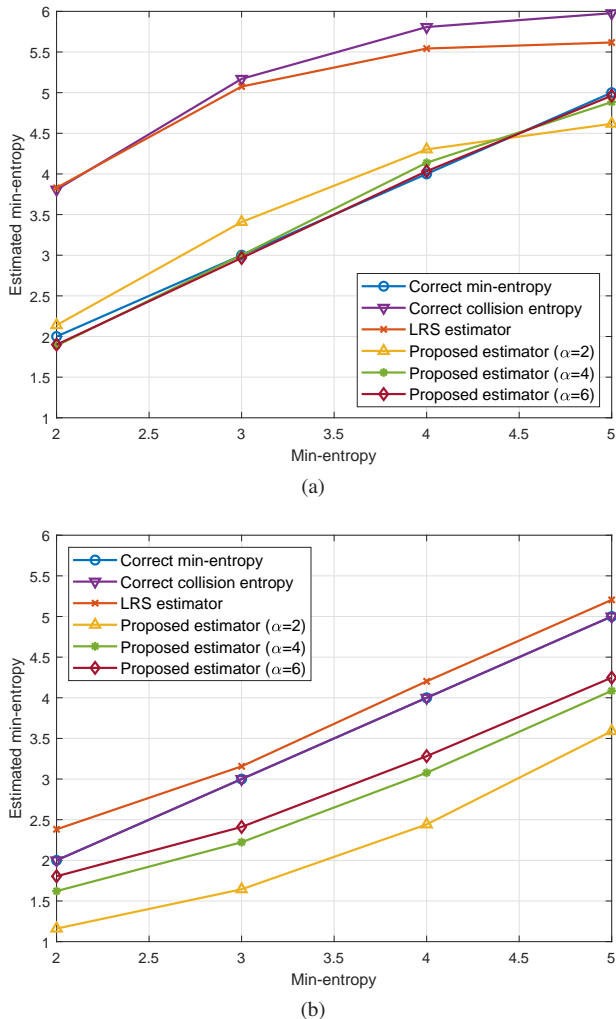
Fig. 5. Comparison of min-entropy estimators for (a) near-uniform distributed sources and (b) inverted near-uniform distributed sources.

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 α.

For the first-order Markov sources, the min-entropy estimators estimate the min-entropy rate. By [17], [18], 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 only the variance of estimates but also the bias as $p \to 0.5$. Note that the improvement of $\mathsf{Var}(\widehat{\theta})$ diminishes as α increases as shown in Fig. 4(b).

Fig. 5 compares the min-entropy estimators for near-uniform distributed sources and inverted near-uniform distributed sources where the alphabet size is 64, i.e., $k = 64$.

	$\alpha = 2$	$\alpha = 4$	$\alpha = 6$
RANDOM.ORG	0.8889	0.9549	0.9572
Ubld.it	0.8277	0.8598	0.8941
LKRNG	0.9364	0.9843	0.9844

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 generalized 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 min-entropy. The proposed estimators underestimate the min-entropy, which can be explained by (15). Since $\widehat{\theta} > \widehat{p}_c$ in (15), we observe that $\widehat{\theta} > \widehat{p}_c \simeq \theta$ for inverted near-uniform distributed sources. Fortunately, the underestimation bias can be reduced by adopting a higher order α 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], [16]. 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 [16, 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.

We also evaluate min-entropy estimates using random number generators deployed in the real-world as in [3], [16]. 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 [19] is a service that provides random numbers based on atmospheric noise and Ubld.it generates random numbers by a TrueRNG device by [20]. The min-entropy estimates of the real-world sources are presented in Table II. We observe that the generalized LRS estimator improves the accuracy of min-entropy estimates by assuming that these real-world sources are high-entropy sources.

We observe that the LRS estimator suffers from significant overestimation problem for most cases of BMS, Markov sources, and near-uniform distributed 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 (improved LRS 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 setting $\alpha = 2$. It was shown that the generalized LRS estimator can improve the bias and variance of min-entropy estimates.

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 α-wise collisions as $D_{\alpha,w}$ for the w-tuples in Step 4 of Algorithm 3, which is given by

$$D_{\alpha,w} = \sum_{i=1}^{k^w} \binom{C_i}{\alpha}, \tag{36}$$

where C_i is the number of occurrences of the ith w-tuple. We suppose that $\binom{C_i}{\alpha} = 0$ 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 α-wise collisions (36).

Lemma 17: For a large L, $\mathbb{E}(D_{\alpha,w+1}) \approx M_\alpha(\mathbf{p}) \cdot \mathbb{E}(D_{\alpha,w})$.

Proof: Denote $\{a_1, \ldots, a_{k^w}\}$ as the alphabet of w-tuples and $\{c_1, \ldots, c_{k^w}\}$ as the number of occurrences of each w-tuple in $\mathbf{s}$. For a w-tuple element a_j for $j \in \{1, \ldots, k^w\}$, 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 α-wise collisions for $(w+1)$-tuples is given by

$$\mathbb{E}(D_{\alpha,w+1}) = \sum_{j=1}^{k^w} \sum_{i=1}^{k} \binom{c_j p_i}{\alpha}$$

$$\approx \sum_{j=1}^{k^w} \sum_{i=1}^{k} \frac{c_j^\alpha p_i^\alpha}{\alpha!} \approx \sum_{j=1}^{k^w} \frac{c_j^\alpha M_\alpha(\mathbf{p})}{\alpha!} \tag{37}$$

$$\approx \sum_{j=1}^{k^w} \binom{c_j}{\alpha} \cdot M_\alpha(\mathbf{p}) \tag{38}$$

$$= M_\alpha(\mathbf{p}) \cdot \mathbb{E}(D_{\alpha,w}), \tag{39}$$

where (37) follows from $\binom{c_j p_i}{\alpha} \approx \frac{c_j^\alpha p_i^\alpha}{\alpha!}$ for $c_j p_i \gg \alpha$ (i.e., for a large L) and Definition 1. If a $c_j p_i$ is not much greater than α, then it can be neglected. Also, (38) follows from $\binom{c_j}{\alpha} \approx \frac{c_j^\alpha}{\alpha!}$. Finally, (39) follows from (36). ∎

For the proof of Theorem 16, we will take the value of v to be $\overline{v}$, 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 α times in the sequence (see Step 2 in Algorithm 3).

Lemma 18: For a large L, $\overline{v} \approx \log_{\frac{1}{M_\alpha}} \binom{l}{\alpha}$.

Proof: By the definition of $\overline{v}$ and (36), $\overline{v}$ is the largest value such that $\mathbb{E}(D_{\alpha,v}) \geq 1$. Hence, $\mathbb{E}(D_{\alpha,1}) \cdot (M_\alpha)^{\overline{v}-1} \geq 1$ and $\mathbb{E}(D_{\alpha,1}) \cdot (M_\alpha)^{\overline{v}} < 1$ by Lemma 17. Then, we can obtain

$$\log_{\frac{1}{M_\alpha}} \mathbb{E}(D_{\alpha,1}) < \overline{v} \leq \log_{\frac{1}{M_\alpha}} \mathbb{E}(D_{\alpha,1}) + 1 \tag{40}$$

$$\log_{\frac{1}{M_\alpha}} \binom{l}{\alpha} M_\alpha < \overline{v} \leq \log_{\frac{1}{M_\alpha}} \binom{l}{\alpha} M_\alpha + 1 \tag{41}$$

$$\log_{\frac{1}{M_\alpha}} \binom{l}{\alpha} - 1 < \overline{v} \leq \log_{\frac{1}{M_\alpha}} \binom{l}{\alpha}. \tag{42}$$

For $l \gg \alpha$, $\overline{v} \approx \log_{\frac{1}{M_\alpha}} \binom{l}{\alpha}$. ∎

Lemma 19: For a uniformly distributed $\mathbf{s} = (s_1, \ldots, s_L)$ with a large L,

$$\frac{\overline{v}_{\alpha+1}}{\overline{v}_\alpha} \approx \frac{\alpha^2 - 1}{\alpha^2}, \tag{43}$$

which is less than one. Note that $\overline{v}_\alpha$ denotes the $\overline{v}$ with order α.

Proof: We can obtain

$$\frac{\overline{v}_{\alpha+1}}{\overline{v}_\alpha} = \frac{\log_k \binom{l_{\alpha+1}}{\alpha+1}}{\alpha} \cdot \frac{\alpha - 1}{\log_k \binom{l_\alpha}{\alpha}} \tag{44}$$

$$= \frac{\alpha - 1}{\alpha} \cdot \frac{\ln \binom{l_{\alpha+1}}{\alpha+1}}{\ln \binom{l_\alpha}{\alpha}} \tag{45}$$

$$\approx \frac{\alpha - 1}{\alpha} \cdot \frac{\ln L^{\alpha+1} - \ln \overline{v}_{\alpha+1}^{\alpha+1} - \ln(\alpha+1)!}{\ln L^\alpha - \ln \overline{v}_\alpha^\alpha - \ln \alpha!} \tag{46}$$

$$\approx \frac{\alpha - 1}{\alpha} \cdot \frac{\ln L^{\alpha+1}}{\ln L^\alpha} \tag{47}$$

$$= \frac{(\alpha - 1)(\alpha + 1)}{\alpha^2}, \tag{48}$$

where (44) follows from Lemma 18 and $M_\alpha = k^{-(\alpha-1)}$ for a uniformly distributed source. For a large L, (46) follows from $\binom{l_\alpha}{\alpha} \approx \frac{l_\alpha^\alpha}{\alpha!}$ and $l_\alpha = \left\lfloor \frac{L}{\overline{v}_\alpha} \right\rfloor \approx \frac{L}{\overline{v}_\alpha}$. Also, (47) follows from $L^\alpha \gg \overline{v}_\alpha^\alpha$ and $L^\alpha \gg \alpha!$ for a large L. ∎

APPENDIX B
PROOF OF THEOREM 16

For every subset $I \subseteq \{1, \ldots, l = \lfloor \frac{L}{w} \rfloor\}$ of size α, we define X_I 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). By (36), it is clear that

$$D_{\alpha,w} = \sum_{|I|=\alpha} X_I \tag{49}$$

and

$$\mathbb{E}(X_I) = M_{\alpha,w}, \tag{50}$$

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 [15].

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,

$$\mathbb{E}(\overline{X}_I \cdot \overline{X}_J) = M_{\alpha+t,w} - M_{\alpha,w}^2 \tag{51}$$

if $|I \cup J| = \alpha + t < 2\alpha$ [15]. Since $M_{\alpha,w} = \frac{1}{k^{w(\alpha-1)}}$ for a uniformly distributed source, we obtain

$$\mathbb{E}(\overline{X}_I \cdot \overline{X}_J) = \frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}}. \tag{52}$$

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

$$\mathsf{Var}(D_{\alpha,w})$$
$$= \sum_{t=0}^{\alpha-1} \sum_{|I \cup J| = \alpha+t} \mathbb{E}(\overline{X}_I \cdot \overline{X}_J) \tag{53}$$

$$= \sum_{t=0}^{\alpha-1} \sum_{|I \cup J| = \alpha+t} \left(\frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}} \right) \tag{54}$$

$$= \sum_{t=0}^{\alpha-1} \binom{l}{\alpha} \binom{l-\alpha}{t} \binom{\alpha}{t} \left(\frac{1}{k^{w(\alpha+t-1)}} - \frac{1}{k^{2w(\alpha-1)}} \right) \tag{55}$$

$$\approx \frac{1}{k^{w(\alpha-1)}} \binom{l}{\alpha} \sum_{t=0}^{\alpha-2} \binom{l}{t} \binom{\alpha}{t} \left(\frac{1}{k^{wt}} - \frac{1}{k^{w(\alpha-1)}} \right), \tag{56}$$

where (54) follows from (52). Also, (56) follows from $\binom{l-\alpha}{t} \approx \binom{l}{t}$ for $l \gg \alpha$ and $\frac{1}{k^{wt}} - \frac{1}{k^{w(\alpha-1)}} = 0$ for $t = \alpha - 1$.

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

$$\mathsf{Var}(\widetilde{M}_{\alpha,w}) = \mathsf{Var}\left(\widehat{M}_{\alpha,w}^{\frac{1}{w}}\right)$$

$$\approx \frac{1}{w^2} \cdot \mathbb{E}(\widehat{M}_{\alpha,w})^{\frac{2(1-w)}{w}} \cdot \mathsf{Var}(\widehat{M}_{\alpha,w}) \tag{57}$$

$$= \frac{1}{w^2} \cdot k^{2(\alpha-1)(w-1)} \cdot \frac{\mathsf{Var}(D_{\alpha,w})}{\binom{l}{\alpha}^2} \tag{58}$$

$$\approx \frac{k^{(\alpha-1)(w-2)}}{w^2} \cdot \frac{\sum_{t=0}^{\alpha-2} \binom{l}{t} \binom{\alpha}{t} \left(k^{-wt} - k^{-w(\alpha-1)} \right)}{\binom{l}{\alpha}}, \tag{59}$$

where (57) follows from the first-order Taylor approximation (i.e., $\mathsf{Var}(f(x)) \approx f'(\mathbb{E}(x))^2 \cdot \mathsf{Var}(x)$ where $f(x) = x^{\frac{1}{w}}$). Since (32) is an unbiased estimator (i.e., $\mathbb{E}(\widehat{M}_{\alpha,w}) = M_{\alpha,w} = k^{-w(\alpha-1)}$), (58) holds. Finally, (59) follows from (56).

Now we show that $\mathsf{Var}(\widetilde{M}_{\alpha,w}) \le \mathsf{Var}(\widetilde{M}_{\alpha,w+1})$ for $w \ge 3$. For each term of (59),

$$\frac{k^{(\alpha-1)(w-2)}}{w^2} \cdot \frac{\binom{l}{t} \binom{\alpha}{t} \left(k^{-wt} - k^{-w(\alpha-1)} \right)}{\binom{l}{\alpha}}$$

$$< \frac{k^{(\alpha-1)(w-2)}}{(w+1)^2} \cdot \frac{k^{\alpha-1}}{k^t} \cdot \frac{\binom{l}{t} \binom{\alpha}{t} \left(k^{-wt} - k^{-w(\alpha-1)} \right)}{\binom{l}{\alpha}} \tag{60}$$

$$= \frac{k^{(\alpha-1)(w-1)}}{(w+1)^2} \cdot \frac{\binom{l}{t} \binom{\alpha}{t} \left(k^{-(w+1)t} - k^{-\{w(\alpha-1)+t\}} \right)}{\binom{l}{\alpha}} \tag{61}$$

$$< \frac{k^{(\alpha-1)(w-1)}}{(w+1)^2} \cdot \frac{\binom{l}{t} \binom{\alpha}{t} \left(k^{-(w+1)t} - k^{-(w+1)(\alpha-1)} \right)}{\binom{l}{\alpha}}, \tag{62}$$

where (60) follows from $\left(\frac{w}{w+1} \right)^2 \cdot \frac{k^{\alpha-1}}{k^t} > 1$ for $k \ge 2$ and $w \ge 3$. Also, (62) follows from $t < \alpha - 1$. Hence,

$$\mathsf{Var}(\widetilde{M}_{\alpha,w}) < \mathsf{Var}(\widetilde{M}_{\alpha,w+1}) \tag{63}$$

for $w \ge 3$.

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 $\mathsf{Var}(\widetilde{M}_\alpha) \approx \mathsf{Var}(\widetilde{M}_{\alpha,\overline{v}})$. Then,

$$\mathsf{Var}(\widetilde{M}_\alpha) \approx \frac{k^{(\alpha-1)(\overline{v}_\alpha-2)}}{\overline{v}_\alpha^2}$$
$$\cdot \frac{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t} \binom{\alpha}{t} \left(k^{-t\overline{v}_\alpha} - k^{-(\alpha-1)\overline{v}_\alpha} \right)}{\binom{l_\alpha}{\alpha}}, \tag{64}$$

where we denote $\overline{v} = \overline{v}_\alpha$ and $l = l_\alpha$ since both $\overline{v}$ and l depend on α.

By the first-order Taylor approximation,

$$\mathsf{Var}(\widehat{\theta}_\alpha) \approx z(\widehat{\theta}_\alpha, \alpha)^2 \cdot \mathsf{Var}(\widetilde{M}_\alpha), \tag{65}$$

where $z(\widehat{\theta}_\alpha, \alpha)$ is given by

$$z(\widehat{\theta}_\alpha, \alpha) = \frac{d\widehat{\theta}_\alpha}{d\widetilde{M}_\alpha} = \frac{1}{\alpha \left\{ \widehat{\theta}_\alpha^{\alpha-1} - \left(\frac{1-\widehat{\theta}_\alpha}{k-1} \right)^{\alpha-1} \right\}}, \tag{66}$$

which is derived from (33).

If $\widetilde{M}_\alpha = \frac{1}{k^{\alpha-1}}$, then $z(\widehat{\theta}_\alpha, \alpha) \to \infty$. However, Algorithm 3 sets $\widehat{\theta}_\alpha = \frac{1}{k}$ for $\widetilde{M}_\alpha \le \frac{1}{k^{\alpha-1}}$ instead of solving (33). Hence, $z(\widehat{\theta}_\alpha, \alpha)$ should be considered only if $\widetilde{M}_\alpha = \frac{1}{k^{\alpha-1}} + \delta$ where $0 < \delta \ll k$ for uniformly distributed sources. Then, we can set $\widehat{\theta}_\alpha = \frac{1}{k} + \delta'$ where $0 < \delta' \ll k$. By [16, Theorem 4], $z(\widehat{\theta}_\alpha, \alpha) \approx \frac{k^{\alpha-3}}{\alpha(\alpha-1)} \cdot \frac{k-1}{\delta'}$. Then,

$$\frac{z(\widehat{\theta}_{\alpha+1}, \alpha+1)}{z(\widehat{\theta}_\alpha, \alpha)} \approx \frac{\alpha-1}{\alpha+1} \cdot k. \tag{67}$$

Then, we obtain

$$\xi(\alpha) = \frac{\mathsf{Var}(\widehat{\theta}_{\alpha+1})}{\mathsf{Var}(\widehat{\theta}_\alpha)}$$

$$\approx \frac{z(\widehat{\theta}_{\alpha+1}, \alpha+1)^2}{z(\widehat{\theta}_\alpha, \alpha)^2} \cdot \frac{\mathsf{Var}(\widetilde{M}_{\alpha+1})}{\mathsf{Var}(\widetilde{M}_\alpha)} \tag{68}$$

$$= \left(\frac{\alpha-1}{\alpha+1} \right)^2 \cdot \left(\frac{\overline{v}_\alpha}{\overline{v}_{\alpha+1}} \right)^2 \cdot \frac{\binom{l_\alpha}{\alpha}}{\binom{l_{\alpha+1}}{\alpha+1}} \cdot k^{\{\alpha\overline{v}_{\alpha+1} - (\alpha-1)\overline{v}_\alpha\}}$$

$$\cdot \frac{\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t} \binom{\alpha+1}{t} \left(k^{-t\overline{v}_{\alpha+1}} - k^{-\alpha\overline{v}_{\alpha+1}} \right)}{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t} \binom{\alpha}{t} \left(k^{-t\overline{v}_\alpha} - k^{-(\alpha-1)\overline{v}_\alpha} \right)} \tag{69}$$

$$\approx \left(\frac{\alpha-1}{\alpha+1} \right)^2 \cdot \left(\frac{\overline{v}_\alpha}{\overline{v}_{\alpha+1}} \right)^2$$

$$\cdot \frac{\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t} \binom{\alpha+1}{t} \left(k^{-t\overline{v}_{\alpha+1}} - k^{-\alpha\overline{v}_{\alpha+1}} \right)}{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t} \binom{\alpha}{t} \left(k^{-t\overline{v}_\alpha} - k^{-(\alpha-1)\overline{v}_\alpha} \right)}, \tag{70}$$

where (68) follows from (65). Also, (69) follows from (64) and (67). By Lemma 18 (see Appendix A) and $M_\alpha = \frac{1}{k^{\alpha-1}}$, we obtain $\overline{v}_\alpha \approx \log_{k^{\alpha-1}} \binom{l_\alpha}{\alpha} = \frac{1}{\alpha-1} \log_k \binom{l_\alpha}{\alpha}$, which leads to

$$k^{\overline{v}_\alpha} \approx \binom{l_\alpha}{\alpha}^{\frac{1}{\alpha-1}}. \tag{71}$$

Then, (70) follows from $k^{\{\alpha\overline{v}_{\alpha+1} - (\alpha-1)\overline{v}_\alpha\}} \approx \frac{\binom{l_{\alpha+1}}{\alpha+1}}{\binom{l_\alpha}{\alpha}}$.

Also, we obtain

$$
\begin{aligned}
\xi(\alpha) &\approx \left(\frac{\alpha}{\alpha+1}\right)^4 \\
&\quad \cdot \frac{\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t}\binom{\alpha+1}{t}\left(k^{-t\overline{v}_{\alpha+1}} - k^{-\alpha\overline{v}_{\alpha+1}}\right)}{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t}\binom{\alpha}{t}\left(k^{-t\overline{v}_\alpha} - k^{-(\alpha-1)\overline{v}_\alpha}\right)}
\end{aligned}
\tag{72}
$$

$$
= \left(\frac{\alpha}{\alpha+1}\right)^4
$$
$$
\cdot \frac{\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t}\binom{\alpha+1}{t}\left\{\binom{l_{\alpha+1}}{\alpha+1}^{-\frac{t}{\alpha}} - \binom{l_{\alpha+1}}{\alpha+1}^{-1}\right\}}{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t}\binom{\alpha}{t}\left\{\binom{l_\alpha}{\alpha}^{-\frac{t}{\alpha-1}} - \binom{l_\alpha}{\alpha}^{-1}\right\}}
\tag{73}
$$

$$
\approx \left(\frac{\alpha}{\alpha+1}\right)^4 \cdot \frac{\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t}\binom{\alpha+1}{t}\binom{l_{\alpha+1}}{\alpha+1}^{-\frac{t}{\alpha}}}{\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t}\binom{\alpha}{t}\binom{l_\alpha}{\alpha}^{-\frac{t}{\alpha-1}}},
\tag{74}
$$

where (72) and (73) follow from Lemma 19 and (71), respectively. Also, (74) follows from

$$
\binom{l_\alpha}{\alpha}^{-\frac{t}{\alpha-1}} \gg \binom{l_\alpha}{\alpha}^{-1}
\tag{75}
$$

for $t \le \alpha - 2$ and a large L. We derive (75) as follows:

$$
\binom{l_\alpha}{\alpha}^{-\frac{t}{\alpha-1}} \ge \binom{l_\alpha}{\alpha}^{-1} \cdot \binom{l_\alpha}{\alpha}^{\frac{1}{\alpha-1}}
\tag{76}
$$

$$
> \binom{l_\alpha}{\alpha}^{-1} \cdot \left(\frac{l_\alpha}{\alpha}\right)^{1+\frac{1}{\alpha-1}}
\tag{77}
$$

$$
\gg \binom{l_\alpha}{\alpha}^{-1},
\tag{78}
$$

where (76) follows from $t \le \alpha - 2$ and (77) follows from $\binom{l_\alpha}{\alpha} > (\frac{l_\alpha}{\alpha})^\alpha$ for $l_\alpha > \alpha$. Also, (78) holds because $l_\alpha \gg \alpha$.

Finally, we show that (74) converges to $\left(\frac{\alpha}{\alpha+1}\right)^4$. For a large L, $l_\alpha = \left\lfloor \frac{L}{\overline{v}_\alpha} \right\rfloor \approx \frac{L}{\overline{v}_\alpha}$ and $\binom{l_\alpha}{t} \approx \frac{l_\alpha^t}{t!}$. Then,

$$
\begin{aligned}
&\sum_{t=0}^{\alpha-1} \binom{l_{\alpha+1}}{t}\binom{\alpha+1}{t}\binom{l_{\alpha+1}}{\alpha+1}^{-\frac{t}{\alpha}} \\
&\approx \sum_{t=0}^{\alpha-1} \binom{\alpha+1}{t} \cdot \frac{(l_{\alpha+1})^{t(1-\frac{\alpha+1}{\alpha})}}{t! \cdot \{(\alpha+1)!\}^{-\frac{t}{\alpha}}}
\end{aligned}
\tag{79}
$$

$$
\approx \sum_{t=0}^{\alpha-1} \binom{\alpha+1}{t} \cdot \frac{\left(\frac{L}{\overline{v}_{\alpha+1}}\right)^{-\frac{t}{\alpha}}}{t! \cdot \{(\alpha+1)!\}^{-\frac{t}{\alpha}}}
\tag{80}
$$

$$
= \sum_{t=0}^{\alpha-1} \binom{\alpha+1}{t} \cdot \frac{\{(\alpha+1)! \cdot \overline{v}_{\alpha+1}\}^{\frac{t}{\alpha}}}{t!} \cdot L^{-\frac{t}{\alpha}}.
\tag{81}
$$

Also,

$$
\begin{aligned}
&\sum_{t=0}^{\alpha-2} \binom{l_\alpha}{t}\binom{\alpha}{t}\binom{l_\alpha}{\alpha}^{-\frac{t}{\alpha-1}} \\
&\approx \sum_{t=0}^{\alpha-2} \binom{\alpha}{t} \cdot \frac{(\alpha! \cdot \overline{v}_\alpha)^{\frac{t}{\alpha-1}}}{t!} \cdot L^{-\frac{t}{\alpha-1}}.
\end{aligned}
\tag{82}
$$

For a large L, (81) converges to one because the highest degree of $L^{-\frac{t}{\alpha}}$ is zero by $t = 0$. Similarly, (82) converges to one. Hence, $\xi(\alpha) \approx \left(\frac{\alpha}{\alpha+1}\right)^4$ for a large L.

References

[1] 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.

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

[3] 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.

[4] T. Amaki, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "A worst-case-aware design methodology for noise-tolerant oscillator-based true random number generator with stochastic behavior modeling," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1331–1342, Aug. 2013.

[5] 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.

[6] 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.

[7] 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.

[8] S. Zhu, Y. Ma, X. Li, J. Yang, J. Lin, and J. Jing, "On the analysis and improvement of min-entropy estimation on time-varying data," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1696–1708, Oct. 2020.

[9] C. Beck and F. Schögl, *Thermodynamics of Chaotic Systems: An Introduction*, ser. Cambridge Nonlinear Science Series. Cambridge University Press, 1993.

[10] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," *Electron. Colloq. Comput. Complexity*, vol. 7, Jan. 2000.

[11] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing closeness of discrete distributions," *J. ACM*, vol. 60, no. 1, pp. 4:1–4:25, Feb. 2013.

[12] 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.

[13] S. Zhu, Y. Ma, T. Chen, J. Lin, and J. Jing, "Analysis and improvement of entropy estimators in NIST SP 800-90B for non-IID entropy sources," *IACR Trans. Symmetric Cryptol.*, vol. 2017, no. 3, pp. 151–168, Sep. 2017.

[14] J. Golic, "On the relationship between the information measures and the Bayes probability of error," *IEEE Trans. Inf. Theory*, vol. 33, no. 5, pp. 681–693, Sep. 1987.

[15] 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

[16] 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.

[17] 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.

[18] 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.

[19] "RANDOM.ORG." [Online]. Available: https://www.random.org

[20] "Ubld.it: TrueRNG." [Online]. Available: http://ubld.it/products/truerng-hardware-random-number-generator/