On the Entropy Bounds for Collision Statistic

Ju-Sung Kang, Yongjin Yeom, and Okyeon Yi

Department of Mathematics,
Cryptography & Information Security Institute,
Kookmin University, Seoul, Korea
{jskang,salt,oyyi}@kookmin.ac.kr

Abstract

Hagerty and Draper [3] presented a theoretical result that explains the probabilistic background of the NIST's entropy test at the random bit generation workshop in 2012. They introduced the notion of an entropic statistic and desired to bound the entropy rate of an unknown output distribution of a given entropy source. However there has been no detailed derivation process about the upper bound, while the theoretical steps of obtaining the lower bound are well described. In this paper we give an elaborate mathematical analysis to obtain the upper bound for the collision statistic. We also present an extended simulation results to investigate practical usefulness of the entropy bounds.

Keywords: *Entropy, Entropic statistic, Collision statistic.*

1 Introduction

An entropy source which generates random bitstrings is fundamentally used in cryptography and many security applications. There has been comprehensive study on the pseudorandom bit generation using deterministic random bit generator (DRBG) and unknown seed value [1]. However the research on creating the unknown value has not been received attention. Recently NIST published a noticeable document about the entropy sources that contains a noise source such as thermal noise or hard drive seek times, used for random bit generation [2]. This publication includes entropy source development requirements and tests for determining entropy provided by entropy sources. In order to estimate an entropy source, NIST uses a very conservative measure known as min-entropy.

On the other hand Hagerty and Draper [3] have presented a theoretical result that explains the probabilistic background of the statistical entropy test of [2] at the Random Bit Generation Workshop 2012 [4]. They suggested the notion of an entropic statistic that corresponds to a constrained optimization problem related to the entropy of some distribution family. By using an appropriate entropic statistic, the authors desired to bound the entropy rate of an unknown output distribution of a given entropy source. The collision statistic is a typical entropic statistic. The upper and lower bounds for the family of probability distributions which have the same expected value of the collision statistic are proposed in [3]. However there has been no detailed derivation process about the upper bound, while the theoretical steps of obtaining the lower bound are well described.

In this paper we show an elaborate mathematical analysis to obtain the upper bound. Our work contributes to solidify a theoretical foundation for the entropy bounds. Our result for the upper bound has not been straightforwardly accomplished by some simply modified arguments from the deriving process for the lower bound. Our approach is slightly different from the original paper, since the used one-parameter family of probability distributions to obtain the upper bound has some implicitly discontinuous points with respect to the parameter. We also provide more generalized simulation results than the original paper in order to investigate practicality of the entropy bounds for the collision statistic.

2 Entropy and Entropic Statistics

Let $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ be a discrete probability distribution on n states $s_1, \ldots, s_n$. The min-entropy of the distribution $\mathbf{p}$ is defined by

$$H_\infty(\mathbf{p}) = \min_{1 \le i \le n} (-\log_2 p_i) = -\log_2 \left(\max_{1 \le i \le n} p_i \right) .$$

We assume that the outputs $X_1, X_2, \ldots, X_N$ from the entropy source are independent identically distributed (i.i.d.) random variables with an unknown probability distribution. In [3], the authors have desired to bound the entropy rate of an unknown distribution given a measurement of an appropriate real-valued statistic by the following optimization problems:

Problem 1. (Minimization)

$$h_S(m) = \min_{\mathbf{p} \in \mathcal{P}_S(m)} H_\infty(\mathbf{p}) , \quad \mathcal{P}_S(m) = \{\mathbf{p} : E_\mathbf{p}[S] = m\} ,$$

where $E_\mathbf{p}[S]$ denotes the expected value of the real-valued statistic S under the probability distribution $\mathbf{p}$.

Problem 2. (Maximization)

$$H_S(m) = \max_{\mathbf{p} \in \mathcal{P}_S(m)} H_\infty(\mathbf{p}) , \quad \mathcal{P}_S(m) = \{\mathbf{p} : E_\mathbf{p}[S] = m\} .$$

Hagerty and Draper [3] presented an efficient procedure for solving these optimization problems for a selected class of statistics which denoted as entropic statistics.

Definition 1 *A real-valued statistic S is said to be entropic with respect to the function H_∞ if for every pair of values m_1 and m_2 such that $m_1 < m_2$ and the sets $\mathcal{P}_S(m_1)$ and $\mathcal{P}_S(m_2)$ are nonempty, $h_S(m_1) < h_S(m_2)$.*

We can obtain the lower and upper bound of min-entropy for a given entropy source with one entropic statistic by solving Problem 1 and Problem 2. Thus it is natural to suggest the following definition.

Definition 2 *Let S be an entropic statistic with respect to the function H_∞. If $-S$ is entropic with respect to the function $-H_\infty$, then we say that S entropically bounds H_∞.*

The following theorem allows us an effective way to solve Problem 1 and Problem 2.

Theorem 1 *[3] Let $\mathbf{p}_\theta$ be an one-parameter family of probability distributions on n states parameterized by $\theta \in [\theta_{min}, \theta_{max}]$. Suppose that for a real-valued statistic S, we have the following properties:*

1. *Monotonicity: $E_{\mathbf{p}_\theta}[S]$ and $H_\infty(\mathbf{p}_\theta)$ are strictly decreasing with respect to the parameter θ. That is,*
$$\frac{d}{d\theta}E_{\mathbf{p}_\theta}[S] < 0 \ , \ \text{and} \ \ \frac{d}{d\theta}H_\infty(\mathbf{p}_\theta) < 0 \ ,$$
for any differentiable point $\theta \in (\theta_{min}, \theta_{max})$.
2. *Convexity: For each $\theta \in [\theta_{min}, \theta_{max}]$, the probability distribution $\mathbf{p}_\theta$ maximizes $E_{\mathbf{p}}[S]$ over all probability distributions having a fixed value of $H_\infty(\mathbf{p}_\theta)$. In other words, we have*
$$E_{\mathbf{p}_\theta}[S] \geq E_{\mathbf{p}}[S] \ , \ \text{for all } \mathbf{p} \text{ such that } H_\infty(\mathbf{p}) = H_\infty(\mathbf{p}_\theta) \ .$$
3. *Surjectivity: For every probability distribution $\mathbf{p}$, there exist values $\theta_1, \theta_2 \in [\theta_{min}, \theta_{max}]$ such that*
$$H_\infty(\mathbf{p}) = H_\infty(\mathbf{p}_{\theta_1}) \ , \quad E_{\mathbf{p}}[S] = E_{\mathbf{p}_{\theta_2}}[S] \ .$$

Then the statistic S is entropic with respect to the function H_∞. Furthermore, there exists a unique value of $\theta \in [\theta_{min}, \theta_{max}]$, such that the probability distribution $\mathbf{p}_\theta$ minimizes H_∞ over the set of distributions having the same expected value of the statistic.

3 Entropy Bounds for Collision Statistic

3.1 Previous Results

The authors of [3] presented the collision statistic as the first example of the entropic statistic. The collision statistic computes the average time up to the first collision of a given output sequence.

Definition 3 *Let X_1, X_2, ... be a sequence of i.i.d. random variables that represent a sequence of outputs from the given entropy source. The collision times T_0, T_1, ... , T_k are defined by $T_0 = 0$ and for $1 \leq i \leq k$, $T_i = \min\{j > T_{i-1} : X_j = X_l \text{ for some } l \in (T_{i-1}, j)\}$. The collision statistic is defined by the average of differences of collision times, that is,*

$$C_k = \frac{1}{k}\sum_{i=1}^{k}(T_i - T_{i-1}) = \frac{T_k}{k} \ .$$

In order to solve two optimization problems with constraints, Problem 1 and Problem 2, by using Theorem 1, we need to define two specific one-parameter families of probability distributions. The one-parameter family for the lower bound is the near-uniform family, and that for the upper bound is the inverted near uniform family.

Definition 4 *The one-parameter family $\{\mathbf{p}_\theta \ : \ \theta \in [0 \ , \ 1]\}$ of probability distributions parameterized by $\theta \in [0, 1]$ on n states, $\{s_1, \ldots, s_n\}$, is said to be the near-uniform family if*

$$\mathbf{p}_\theta(s_1) = \theta \ \ \text{and} \ \ \mathbf{p}_\theta(s_i) = \frac{1-\theta}{n-1} \ , \ \text{for } 2 \leq i \leq n \ ,$$

where for each $1 \leq i \leq n$, $\mathbf{p}_\theta(s_i)$ denotes the probability mass on the state s_i.

Definition 5 *The one-parameter family* $\{\mathbf{p}_\psi : \psi \in [0\ ,\ 1]\}$ *of probability distributions parameterized by* $\psi \in [0,1]$ *on* n *states,* $\{s_1, \ldots, s_n\}$*, is said to be the inverted near-uniform family if*

$$\begin{aligned} \mathbf{p}_\psi(s_i) &= \psi \ \text{ for } 1 \le i \le \lfloor \frac{1}{\psi} \rfloor\ , \\ \mathbf{p}_\psi(s_{\lfloor \frac{1}{\psi} \rfloor + 1}) &= 1 - \lfloor \frac{1}{\psi} \rfloor \psi\ , \ \text{and} \\ \mathbf{p}_\psi(s_i) &= 0 \ \text{ for } \lfloor \frac{1}{\psi} \rfloor + 2 \le i \le n\ , \end{aligned}$$

where for each $1 \le i \le n$, $\mathbf{p}_\psi(s_i)$ *denotes the probability mass on the state* s_i *and* $\lfloor x \rfloor$ *is the largest integer not greater than* x*.*

Applying Theorem 1, the authors of [3] proved that the collision statistic C_k is entropic with respect to H_∞ with the minimal distribution being a near-uniform distribution. On the other hand, there has been no detail and concrete derivation about the upper bound in the following theorem. We assume that the number of collision times k is fixed, and hereafter write the statistic C_k simply as C.

Theorem 2 *Let* $\mathbf{p}$ *be a probability distribution on* n *states and let* C *be the collision statistic. Let us choose* $\theta, \psi \in [1/n\ ,\ 1]$ *such that* $\mathbf{p}_\theta$ *is a near-uniform distribution with* $E_{\mathbf{p}_\theta}[C] = E_{\mathbf{p}}[C]$*, and* $\mathbf{p}_\psi$ *is an inverted near-uniform distribution with* $E_{\mathbf{p}_\psi}[C] = E_{\mathbf{p}}[C]$*. Then the min-entropy of* $\mathbf{p}$ *is bounded below and above by* $H_\infty(\mathbf{p}_\theta)$ *and* $H_\infty(\mathbf{p}_\psi)$*, respectively. That is,*

$$H_\infty(\mathbf{p}_\theta) \le H_\infty(\mathbf{p}) \le H_\infty(\mathbf{p}_\psi)\ .$$

3.2 Our Results for the Upper Bound

In this subsection we show that the result for the upper bound in Theorem 2 has not been straightforwardly accomplished by the same arguments for the lower bound. Our approach is slightly different from the original paper, since the used one-parameter family $\{\mathbf{p}_\psi : \psi \in [1/n\ ,\ 1]\}$ of inverted near-uniform distributions to obtain the upper bound has some implicitly discontinuous points with respect to the parameter ψ. In fact, the function $\xi = \mathbf{p}_\psi(s_{\lfloor \frac{1}{\psi} \rfloor + 1}) = 1 - \lfloor \frac{1}{\psi} \rfloor \psi$ is discontinuous at all points of the form $\frac{1}{m}$, where $m = 1, 2, \ldots, n$. We also make use of Theorem 1, but the detailed processes of proving the theorem are quite different from that of the original paper.

Theorem 3 *The collision statistic* $-C$ *is entropic with respect to* $-H_\infty$ *with the optimal distribution being an inverted near-uniform distribution.*

By Theorem 1, we have to show the monotonicity, the convexity, and the surjectivity. So we split the proof of Theorem 3 into several lemmas. At first, we show that in spite of the discontinuous property of $\xi = \mathbf{p}_\psi(s_{\lfloor \frac{1}{\psi} \rfloor + 1}) = 1 - \lfloor \frac{1}{\psi} \rfloor \psi$, $H_\infty(\mathbf{p}_\psi)$ and $E_{\mathbf{p}_\psi}[C]$ are continuous with respect to $\psi \in [1/n\ ,\ 1]$.

Lemma 1 *Let* $\{\mathbf{p}_\psi : \psi \in [0,1]\}$ *be the inverted near-uniform family on* n *states* $\{s_1, \ldots, s_n\}$ *and* C *be the collision statistic. Then* $H_\infty(\mathbf{p}_\psi)$ *and* $E_{\mathbf{p}_\psi}[C]$ *are continuous with respect to the parameter* $\psi \in [1/n\ ,\ 1]$*.*

Proof. Note that $H_\infty(\mathbf{p}_\psi) = -\log_2 \psi$ is continuous for $0 < \psi \leq 1$. Thus $H_\infty(\mathbf{p}_\psi)$ is also continuous on $\psi \in [1/n\ ,\ 1]$.

Now we prove the continuity of the function $f(\psi) = E_{\mathbf{p}_\psi}[C]$. If $\psi = \frac{1}{m}$ for some integer $m = 1, 2, \ldots, n$, then for each $i \leq m$,

$$Pr\,(T_1 > i) = m(m-1)\cdots(m-i+1)\cdot\psi^i\ .$$

Hence, we obtain

$$E_{\mathbf{p}_\psi}[C] = E_{\mathbf{p}_\psi}[T_1] = \sum_{i=1}^{m} Pr\,(T_1 > i) = \sum_{i=1}^{m} \frac{m!}{(m-i)!}\psi^i\ .$$

If $\frac{1}{m+1} < \psi < \frac{1}{m}$ for some integer $m = 1, 2, \ldots, n-1$, then

$$\xi = \mathbf{p}_\psi(s_{\lfloor \frac{1}{\psi}\rfloor+1}) = \mathbf{p}_\psi(s_{m+1}) = 1 - m\psi$$

and for all $i \geq m+2$, $Pr\,(T_1 > i) = 0$. For each $i \leq m+1$,

$$Pr\,(T_1 > i) = i!\left(\binom{m}{i-1}\psi^{i-1}\xi + \binom{m}{i}\psi^i\right)\ .$$

Thus

$$\begin{aligned} E_{\mathbf{p}_\psi}[C_k] &= E_{\mathbf{p}_\psi}[T_1] = \sum_{i=1}^{m+1} Pr\,(T_1 > i) \\ &= \sum_{i=1}^{m+1}\left(\frac{i\cdot m!}{(m-i+1)!}\psi^{i-1}(1-m\psi) + \frac{m!}{(m-i)!}\psi^i\right)\ . \end{aligned}$$

In order to prove the continuity of the function $f(\psi) = E_{\mathbf{p}_\psi}[C]$ on $\psi \in [1/n\ ,\ 1]$, it suffices to show that for any integer $m = 1, 2, \ldots, n$, $\lim_{\psi\to\frac{1}{m}} f(\psi) = f(\frac{1}{m})$, since $f(\psi)$ is represented by a polynomial. At first we show that $f(\psi)$ is left-continuous at $\psi = \frac{1}{m}$.

$$\begin{aligned} \lim_{\psi\to\frac{1}{m}-} f(\psi) &= \lim_{\psi\to\frac{1}{m}-}\sum_{i=1}^{m+1}\left(\frac{i\cdot m!}{(m-i+1)!}\psi^{i-1}(1-m\psi) + \frac{m!}{(m-i)!}\psi^i\right) \\ &= \sum_{i=1}^{m}\frac{i\cdot m!}{(m-i+1)!}\lim_{\psi\to\frac{1}{m}-}\psi^{i-1}(1-m\psi) + \sum_{i=1}^{m}\frac{m!}{(m-i)!}\lim_{\psi\to\frac{1}{m}-}\psi^i \\ &= 0 + \sum_{i=1}^{m}\frac{m!}{(m-i)!}\left(\frac{1}{m}\right)^i = f(1/m)\ . \end{aligned}$$

Secondly we show that $f(\psi)$ is also right-continuous at $\psi = \frac{1}{m}$. Without loss of generality we can assume that $\frac{1}{m} < \psi < \frac{1}{m-1}$, so in this case,

$$\xi = \mathbf{p}_\psi(s_{\lfloor \frac{1}{\psi}\rfloor+1}) = \mathbf{p}_\psi(s_m) = 1 - (m-1)\psi\ .$$

Thus

$$\begin{aligned}
\lim_{\psi\to\frac{1}{m}+} f(\psi) &= \lim_{\psi\to\frac{1}{m}+} \sum_{i=1}^{m} \left(\frac{i\cdot(m-1)!}{(m-i)!}\psi^{i-1}(1-(m-1)\psi) + \frac{(m-1)!}{(m-i-1)!}\psi^{i} \right) \\
&= \sum_{i=1}^{m-1} \frac{i\cdot(m-1)!}{(m-i)!} \lim_{\psi\to\frac{1}{m}+} \psi^{i-1}(1-(m-1)\psi) \\
&\quad + \sum_{i=1}^{m-1} \frac{(m-1)!}{(m-i-1)!} \lim_{\psi\to\frac{1}{m}+} \psi^{i} + \lim_{\psi\to\frac{1}{m}+} m!\psi^{m-1}(1-(m-1)\psi) \\
&= \sum_{i=1}^{m-1} \frac{i\cdot(m-1)!}{(m-i)!} \left(\frac{1}{m}\right)^{i-1} \left(1-\frac{m-1}{m}\right) \\
&\quad + \sum_{i=1}^{m-1} \frac{(m-1)!}{(m-i-1)!} \left(\frac{1}{m}\right)^{i} + m! \left(\frac{1}{m}\right)^{m-1} \left(1-\frac{m-1}{m}\right) \\
&= \sum_{i=1}^{m-1} \left\{ \frac{i\cdot(m-1)!}{(m-i)!} \left(\frac{1}{m}\right)^{i} + \frac{(m-1)!}{(m-i-1)!} \left(\frac{1}{m}\right)^{i} \right\} + m! \left(\frac{1}{m}\right)^{m} \\
&= \sum_{i=1}^{m} \frac{m!}{(m-i)!} \left(\frac{1}{m}\right)^{i} = f(1/m) \, .
\end{aligned}$$

Consequently we obtain that $f(\psi) = E_{\mathbf{p}_\psi}[C]$ is a continuous function on $\psi \in [1/n\,,\,1]$. □

By using the above lemma, we can show that the negative collision statistic $-C$ and the negative min-entropy $-H_\infty$ satisfy the monotonicity and surjectivity conditions of Theorem 1. However we need to modify the parameter $\psi \in [1/n\,,\,1]$ as the parameter $\zeta = 1-\psi$. Since there is an exact one-to-one correspondence between ψ and ζ, two one-parameter families of probability distributions parametrized by ψ and ζ also have an exact one-to-one correspondence. In other words,

$$\{\mathbf{p}_\psi : \psi \in [1/n\,,\,1]\} = \{\mathbf{p}_\zeta \,:\, \zeta \in [0\,,\,1-1/n]\}.$$

Lemma 2 (Monotonicity) *$E_{\mathbf{p}_\zeta}[-C]$ and $-H_\infty(\mathbf{p}_\zeta)$ are strictly decreasing with respect to the parameter $\zeta \in [0\,,\,1-1/n]$.*

Proof. Since $-H_\infty(\mathbf{p}_\zeta) = \log_2(1-\zeta)$, we obtain that

$$\frac{d}{d\zeta}\left(-H_\infty(\mathbf{p}_\zeta)\right) = -\frac{1}{1-\zeta} < 0 \text{ for } \zeta \in (0\,,\,1-1/n) \, .$$

This shows the monotonicity of $-H_\infty(\mathbf{p}_\zeta)$ with respect to the parameter ζ.

In order to prove the monotonicity of the statistic $-C$, we have to show that

$$\frac{d}{d\zeta} E_{\mathbf{p}_\zeta}[-C] < 0 \text{ for } \zeta \in (0\,,\,1-1/n).$$

This inequality can be equivalently written by

$$\frac{d}{d\zeta} E_{\mathbf{p}_\zeta}[C] > 0 \iff \frac{d}{d\psi} E_{\mathbf{p}_\psi}[C] < 0 \, ,$$

Since $\zeta = 1 - \psi$, it suffices to show that

$$\frac{d}{d\psi} E_{\mathbf{p}_\psi}[C] < 0 \text{ for } \psi \in (1/n\,,\, 1).$$

From the proof of Lemma 1, we already have that for any $\psi \in (\frac{1}{m+1}, \frac{1}{m})$,

$$\begin{aligned} E_{\mathbf{p}_\psi}[C] &= \sum_{i=1}^{m+1} g_i(\psi) \\ &= \sum_{i=1}^{m} \left(\frac{i \cdot m!}{(m-i+1)!} \psi^{i-1}(1-m\psi) + \frac{m!}{(m-i)!}\psi^i \right) + (m+1)!\psi^m(1-m\psi)\,. \end{aligned}$$

By defferentiating the summands $g_i(\psi)$ of this formula, we obtain that

$$\begin{aligned} \frac{d}{d\psi} g_{m+1}(\psi) &= (m+1)!\left(m\psi^{m-1}(1-m\psi) + \psi^m(-m)\right) \\ &= (m+1)!m\psi^{m-1}(1-(m+1)\psi) \\ &< 0\,, \quad \text{since } \frac{1}{m+1} < \psi < \frac{1}{m} \end{aligned}$$

and for any $1 \le i \le m$,

$$\begin{aligned} \frac{d}{d\psi} g_i(\psi) &= \frac{i \cdot m!}{(m-i+1)!}\left((i-1)\psi^{i-2}(1-m\psi) + \psi^{i-1}(-m)\right) + \frac{m!}{(m-i)!} i\psi^{i-1} \\ &= \frac{i \cdot m!}{(m-i)!}\psi^{i-2}\left(\frac{i-1}{m-i+1} + \psi(1 - \frac{im}{m-i+1})\right) \\ &= \frac{i \cdot m!}{(m-i+1)!}\psi^{i-2}(i-1)\,(1-(m+1)\psi) \\ &< 0\,. \end{aligned}$$

Therefore for any $1 \le i \le m+1$, $\frac{d}{d\psi} g_i(\psi)$ has a negative value, where $\frac{1}{m+1} < \psi < \frac{1}{m}$ and $m = 1, 2, \ldots, n-1$. This implies that for any two points ψ_1 and ψ_2 in the same sub-interval $(\frac{1}{m+1}, \frac{1}{m})$,

$$E_{\mathbf{p}_{\psi_1}}[C] > E_{\mathbf{p}_{\psi_2}}[C] \;\text{ if } \psi_1 < \psi_2\,.$$

However for any two points ψ_1 and ψ_2 lie in two different sub-intervals, we also obtain the same argument. In fact, by using the continuity of Lemma 1, we obtain that $E_{\mathbf{p}_{\psi_1}}[C] > E_{\mathbf{p}_{\psi_2}}[C]$ for $\frac{1}{m+2} < \psi_1 < \frac{1}{m+1} < \psi_2 < \frac{1}{m}$. This completes the fact that $E_{\mathbf{p}_\psi}[C]$ is a monotonically decreasing function with respect to the parameter ψ. □

Combining these lemmas with the convexity and surjectivity of one parameter family of probability distributions $\mathbf{p}_\psi$, we obtain the upper bound in Theorem 2 which can be generated by the inverted near-uniform distribution.

4 Experimental Results

Suppose that a data set $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ is given and its entropy is about to be measured where x_i is a block which has n possible states. For a given $t \in [1/n, 1]$, the

expected value of collision statistic for $\mathcal{D}$ is bounded below and above by those of near-uniform distribution $\mathbf{p}_{\theta=t}$ and inverted near-uniform distribution $\mathbf{p}_{\psi=t}$, respectively. Note that the parameter t means the maximum probability among n states.

When a data set is provided as a sequence of nibble(4-bit) blocks, the number of possible states n is 16. The expected value of collision statistic is contained in the region enclosed two curves generated by $\mathbf{p}_\theta$ and $\mathbf{p}_\psi$ as Figure 1.

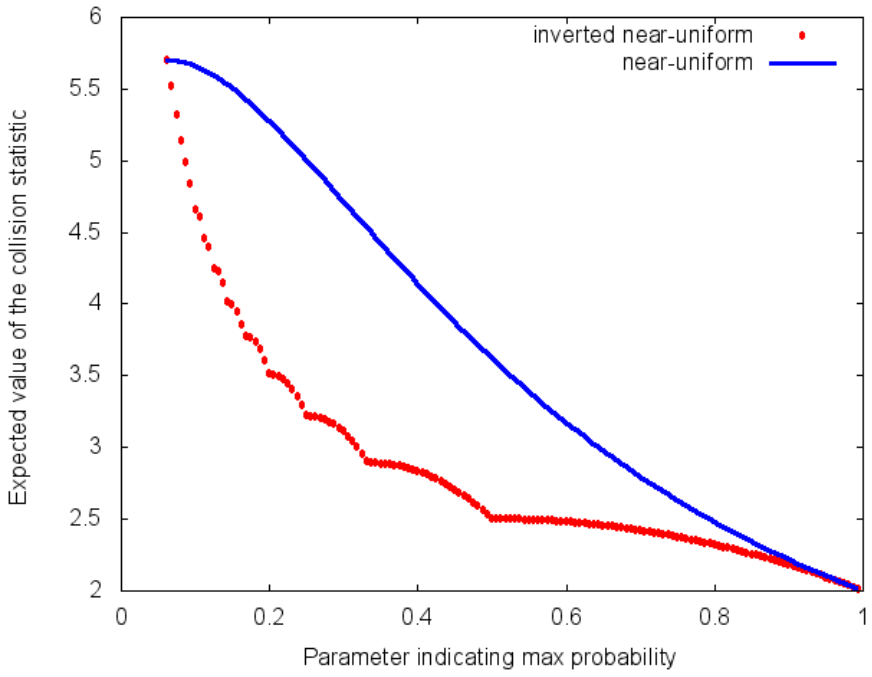

Figure 1. Entropy bounds for collision statistic on 4-bit blocks (16 states)

After calculating the collision statistic $C_{observed}$ for a data set, we can find corresponding θ_0 and ψ_0 satisfying $E_{\mathbf{p}_{\theta=\theta_0}} = C_{observed} = E_{\mathbf{p}_{\psi=\psi_0}}$. If the horizontal line $y = C_{observed}$ intersects two curves $y = E_{\mathbf{p}_\theta}$ and $y = E_{\mathbf{p}_\psi}$ at θ_0 and ψ_0, respectively, the min-entropy $H_\infty(\mathcal{D})$ is expected to be bounded as $-\log_2(\theta_0) \leq H_\infty(\mathcal{D}) \leq -\log_2(\psi_0)$.

Table 1 estimates min-entropy of several data sets (pseudo-random number, output of LFSR, pseudo-random number with parity bits). In the 3rd and the 4th columns, we use pseudo random number except for the last bit (parity bit) of each sample so that the estimations of their min-entropy are bounded by 3 and 7, respectively. With the limited amount of data, we adopt the confidence interval of 95% so that entropy bound is conservatively estimated with confidence level. In the last row of Table 1 min-entropy is optimistically estimated by direct counting of frequency for each state according to the definition of min-entropy.

The entropy estimated by the collision statistic in Table 1 looks very conservertive. Thus, it is generally expected that min-entropy is actually larger than estimated. However, it can be sometimes overestimated for a data set which is far from i.i.d. For example, output of LFSR is estimated as almost full entropy even though its entropy is bounded by the size of LFSR. For that reason, the min-entropy estimation can be effective only for data from almost i.i.d.

Table 1. Estimations of min-entropy of 4-bit blocks for serveral data sets

estimation of min-entropy	pseudo-random	LFSR (8 bits)	4 bit random (even parity)	8 bit random (even parity)
using $C_{observed}$	3.32	3.84	1.41	2.48
with confidence level (95%)	3.04	3.50	1.38	2.37
frequency counting	3.92	3.99	2.98	3.90

5 Conclusion

We have surveyed the paper of Hagerty and Draper [3] that have presented a theoretical result that supports the NIST's entropy test. By suggesting the notion of an entropic statistic and by using an appropriate entropic statistic, the authors desired to bound the entropy rate of an unknown output distribution of a given entropy source. However there has been no detailed derivation process about the upper bound, while the theoretical steps of obtaining the lower bound are well described. In this paper we have presented an elaborate mathematical analysis to obtain the upper bound for the collision statistic. We also have shown more generalized simulation results than the original paper in order to investigate practicality of the entropy bounds.

Acknowledgement

This work is a part of the results of the research "Development of the wide-band underwater mobile communication systems" supported by Ministry of Land, Transportation and Maritime Affairs, Korea.

References

[1] NIST Special Publication 800-90A, *Recommendation for Random Number Generation Using Deterministic Random Bit Generators*, January 2012.

[2] NIST DRAFT Special Publication 800-90B, *Recommendation for the Entropy Sources Used for Random Bit Generation*, August 2012.

[3] P. Hagerty and T. Draper, *Entropy Bounds and Statistical Tests*, Random Bite Generation Workshop 2012, December 2012.

[4] NIST, Random Bit Generation Workshop, `http://www.nist.gov/itl/csd/ct/rbg_workshop2012.cfm`, December 2012.