ENTROPY BOUNDS AND STATISTICAL TESTS

Patrick Hagerty Tom Draper

Abstract

We convert a generic class of entropy tests from pass/fail to a measure of entropy. The conversion enables one to specify a fundamental design criterion: state the number of outputs from a noise source required to satisfy a security threshold. We define new entropy measurements based on a three-step strategy: 1) compute a statistic on raw output of a noise source, 2) define a set of probability distributions based on the result, and 3) minimize the entropy over the set. We present an efficient algorithm for solving the minimization problem for a select class of statistics, denoted as "entropic" statistics; we include several detailed examples of entropic statistics. Advantages of entropic statistics over previous entropy tests include the ability to rigorously bound the entropy estimate, a reduced data requirement, and partial entropy credit for sources lacking full entropy.

1. Good Entropy Sources and Tests

Consider a four-sided die (with faces labeled 1,2,3,4) that is weighted so that the probability of the value i's being rolled is p_i. If each p_i has the value of $\frac{1}{4}$, then we consider the die a fair die; we have no problem using the fair die as a source of randomness with each roll producing two bits of randomness. If the values of p_i are not equal, then the die is considered weighted. There are a number of questions about the weighted die's output that arise.

Question 1.1. *Can the weighted die be used as a good source of randomness?*

Question 1.2. *How many rolls of the die are required to generate n random bits?*

Question 1.3. *How do the answers to the previous questions differ if the values of p_i are unknown versus if the values of p_i are known?*

Question 1.4. *What complications arise if one considers a die with a large number of sides versus a die with four sides?*

The concept that a source can lack full entropy and nevertheless be an excellent source of randomness has eluded some designers of random sources. There are suites of statistical tests that test for full entropy in a source: the source fails the test if the probability is not uniformly distributed among the possible output states. In particular, the weighted die would fail the statisical tests. One goal of this paper is to convert the pass/fail tests into a measure of entropy: the weighted die would receive partial credit as a source of randomness. With

the proper understanding of the entropy produced by the source one can answer Question 1.2, however the actual generation process is beyond the scope of the paper.

The simple approach to Question 1.2 is to generate enough output to model the p_i then compute the entropy of the modeled distribution. Three complications that often arise with this approach: the amount of output may be limited, outputs may be dependent, and there are many notions of entropy. In section 2, we discuss the notions of entropy. When data is limited, it may be impossible to estimate a probabilty distribution for the data; however, one may be able to compute a derived value that is used to estimate the entropy of the outputs. Most of this paper addresses issues related to this computation including:

1. What is the best statistic to compute from a random source?
2. How does one go from a statistic to a measure of randomness?
3. How does one combine multiple statistics to refine an estimate of entropy?

The most difficult complication to address is dependency in outputs—we attempt to address this difficulty by admitting a simple dependent model of the data and performing analysis based on this model. It may be impractical or impossible to obtain an accurate model of the dependence between outputs; thus, we limit our scope of dependence to Markov processes.

2. Rényi Entropy

There are many measures of disorder, or entropy. In [1], Rényi developed a class of entropy functions parameterized by a value $\alpha \in [1, \infty]$:

$$H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right), \tag{2.1}$$

where the probability distribution $\mathbf{p}$ on n states has the probability of p_k for states $k = 1, \cdots, n$. In the limiting case of H_∞, or min-entropy, we employ L'Hôpital's Rule to obtain:

$$H_\infty(\mathbf{p}) = \min_{i=1,\cdots,n} (-\log_2 p_i). \tag{2.2}$$

Shannon entropy is a natural member of this class (also employing L'Hôpital's Rule):

$$H_1(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{2.3}$$

For evaluation purposes, we are interested in min-entropy, H_∞ and Shannon entropy, H_1. We associate a cost of guessing the output of a noise source with the entropy of the noise source. Min-entropy is a metric that measures the difficulty of guessing the easiest to guess output of noise source. Shannon entropy is a metric that measures the difficulty guessing a typical output of a noise source. The results in this paper attempt to bound H_∞ based on some data sampling. For *iid* sources, the lack of dependence simplifies the computations. When one has dependence, the entropy does not scale linearly with the amount of output. In the dependent case, one has to look at the entire block of outputs as one state and compute the entropy over all possible blocks.

3. Entropic Statistics

In this section we develop the theory on which examples in section 4 are based. Our main result states criteria that a statistic must satisfy in order to relate a statistical test to an entropy bound. We assume in this section that the output from the source are *iid* random variables; further, we assume that the probability distribution of the output is unknown. In the *iid* setting, the notion of entropy rate, or entropy per output, is well defined. Most of the examples will choose H_∞ as the entropy function, but the theory is general enough to apply to other measures of entropy. It is common for an evaluator of a noise source to have a sequence of its outputs yet not be able to resolve the probability distribution of the outputs. While the full distribution is required to compute the entropy, one may be able to state upper and lower bounds on the entropy with a prescribed degree of certainty by looking at the data. Explicitly, we desire to bound the entropy rate of an unknown distribution given a measurement, m, of a real-valued statistic, S, by optimizing over all probability distributions having expected value equal to the observed measurement, m.

Problem 3.1. *Solve the constrained optimization problem:*

$$h_S(m) = \min_{\mathbf{p} \in \mathcal{P}_S(m)} H(\mathbf{p}), \tag{3.1}$$

where

$$\mathcal{P}_S(m) = \{\mathbf{p} : \mathbb{E}_{\mathbf{p}}[S] = m\}. \tag{3.2}$$

The notation $\mathbb{E}_{\mathbf{p}}[S]$ is the expected value of the real-valued statistic S under the probability distribution $\mathbf{p}$.

To accommodate noisy measurements and confidence intervals, we often perform the minimization over a superset of $\mathcal{P}_S(m)$. Furthermore, we desire that $h_S(m)$ be monotonic with respect to m, suggesting the notion of being "entropic."

Definition 3.2. We say that a real-valued statistic, S, is **entropic** with respect to the function H if for every pair of values, m and m', such that the sets $\mathcal{P}_S(m)$ and $\mathcal{P}_S(m')$ are non-empty and $m' < m$, we have $h_S(m') < h_S(m)$.

We must solve the potentially difficult minimization problem with constraints before the notion of an entropic statistic is useful. The following theorem allows us effectively to interchange the constraints and the function to be optimized under certain criteria, converting a difficult minimization problem to an easier one (usually by simplifying the constraints).

Theorem 3.1 (Entropic)**.** *Let $\mathbf{p}_\theta$ be a one-parameter family of probability distributions on n states parameterized by $\theta \in [\theta_{min}, \theta_{max}]$. Suppose we have the following properties:*

1. **Monotonicity.** *The expected value of the real-valued statistic, S, is differentiable with respect to the parameter θ and is strictly decreasing. Also, the function H is strictly*

decreasing with respect to the parameter θ. In other words, we have

$$\begin{aligned} \frac{d}{d\theta}\mathbb{E}_{\mathbf{p}_\theta}[S] &< 0, \\ \frac{d}{d\theta}H(\mathbf{p}_\theta) &< 0, \end{aligned}$$

for $\theta \in (\theta_{min}, \theta_{max})$.

2. **Convexity.** *For each $\theta \in [\theta_{min}, \theta_{max}]$, the probability distribution $\mathbf{p}_\theta$ maximizes $\mathbb{E}_{\mathbf{p}_\theta}[S]$ over all probability distributions having a fixed value of $H(\mathbf{p}_\theta)$. Explicitly, we have*

$$\mathbb{E}_{\mathbf{p}_\theta}[S] \geq \mathbb{E}_{\mathbf{p}}[S], \tag{3.3}$$

for all $\mathbf{p}$ such that $H(\mathbf{p}) = H(\mathbf{p}_\theta)$.

3. **Surjectivity.** *For every probability distribution $\mathbf{p}$, there exist values $\theta, \theta' \in [\theta_{min}, \theta_{max}]$ such that*

$$\begin{aligned} H(\mathbf{p}) &= H(\mathbf{p}_\theta), \\ \mathbb{E}_{\mathbf{p}}[S] &= \mathbb{E}_{\mathbf{p}_{\theta'}}[S]. \end{aligned}$$

By monotonicity the values are unique but not necessarily distinct.

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.

Proof. Let S be a real-valued statistic and suppose we have a one-parameter family of probability distributions $\mathbf{p}_\theta$ that satisfies the monotonicity, convexity, and surjectivity properties with respect to the statistic S and function H.

We first show that there is a value of θ such that $\mathbf{p}_\theta$ minimizes H over all probability distributions with a fixed expected value of the statistic S. Let $\mathcal{G} = \{\mathbf{p} : \mathbb{E}_{\mathbf{p}}[S] = m\}$ for a fixed value of m. By surjectivity, there exists $\theta \in [\theta_{\min}, \theta_{\max}]$ such that

$$\mathbb{E}_{\mathbf{g}}[S] = \mathbb{E}_{\mathbf{p}_\theta}[S], \tag{3.4}$$

for all $\mathbf{g} \in \mathcal{G}$. Now fix $\mathbf{g} \in \mathcal{G}$. By surjectivity, there exists a $\theta_{\mathbf{g}} \in [\theta_{\min}, \theta_{\max}]$ such that

$$H\left(\mathbf{p}_{\theta_{\mathbf{g}}}\right) = H(\mathbf{g}). \tag{3.5}$$

By convexity, we have

$$\mathbb{E}_{\mathbf{p}_{\theta_{\mathbf{g}}}}[S] \geq \mathbb{E}_{\mathbf{g}}[S] = \mathbb{E}_{\mathbf{p}_\theta}[S]. \tag{3.6}$$

By monotonicity, we have $\theta_{\mathbf{g}} \leq \theta$. Also by monotonicity, we have

$$H(\mathbf{g}) = H\left(\mathbf{p}_{\theta_{\mathbf{g}}}\right) \geq H(\mathbf{p}_\theta). \tag{3.7}$$

Thus, for each $\mathbf{g} \in \mathcal{G}$, we have

$$H(\mathbf{g}) \geq H\left(\mathbf{p}_\theta\right). \tag{3.8}$$

We now desire to show that S is entropic with respect to H. Let m and m' be distinct values such that $\mathcal{P}_S(m)$ and $\mathcal{P}_S(m')$ are non-empty and $m' < m$. By the preceding argument, there exists $\theta, \theta' \in [\theta_{\min}, \theta_{\max}]$ such that

$$\begin{aligned} H\left(\mathbf{p}_\theta\right) &= \min_{\mathbf{p} \in \mathcal{P}_S(m)} H(\mathbf{p}), \\ H\left(\mathbf{p}_{\theta'}\right) &= \min_{\mathbf{p} \in \mathcal{P}_S(m')} H(\mathbf{p}). \end{aligned}$$

By monotonicity of $\mathbb{E}[S]$, we have $\theta < \theta'$. Subsequently, the monotonicity of H yields

$$H\left(\mathbf{p}_{\theta'}\right) < H\left(\mathbf{p}_\theta\right). \tag{3.9}$$

This completes the proof. □

There are times that one desires an upper bound for the entropy estimate. One method for finding an upper bound is to replace the entropy function, H, with its negative, show that the negative of the statistic is entropic with respect to $-H$ and solve Problem 3.1. Essentially, this is equivalent to solving the following problem:

Problem 3.3. *Solve the constrained optimization problem:*

$$H_S(m) = \max_{\mathbf{p} \in \mathcal{P}_S(m)} H\left(\mathbf{p}\right), \tag{3.10}$$

where

$$\mathcal{P}_S(m) = \left\{\mathbf{p} : \mathbb{E}_{\mathbf{p}}[S] = m\right\}. \tag{3.11}$$

Solving Problems 3.1 and 3.3 allows us to bound the entropy above and below with one statistic, suggesting the following definition.

Definition 3.4. Let S be a real-valued statistic that is entropic with respect to a function H. If $-S$ in entropic with respect to $-H$, then we say that the S **entropically bounds** H.

4. Entropic Examples

In this section, we present examples of statistics and bound entropy measurements based on the statistics. For each of the statistics presented, we solve Problems 3.1 and 3.3. Before we present the statistics, we discuss four selection criteria for the statistics.

The paramount criterion is that the statistic should be related to the entropy of the noise source–entropic statistics. We will apply the tools in the previous section to prove rigorous entropy bounds.

data set	ABCAACBCABBCABACBCBCA
Collision Blocks	ABCA ACBC ABB CABA CBC
Collision Repeat Rate	4 4 3 4 3
Compression Blocks A	A..AA...A...A.A.....A
Compression Repeat Rate A	1..31...4...4.2.....6
Compression Blocks B	.B....B..BB..B..B.B..
Compression Repeat Rate B	.2....5..31..3..3.2..
Compression Blocks C	..C..C.C...C...C.C.C.
Compression Repeat Rate C	..3..3.2...4...4.2.2.
Compression Repeat Rate	123313524314432432226
Partial Collection Blocks	ABC AAC BCA BBC ABA CBC BCA
Partial Collection Repeat Rate	0 1 0 1 1 1 0
Partial Collection Distinct Elements	3 2 3 2 2 2 3

Table 1: Example of repeat rate information for collision, compression, partial collection. The corresponding statistics uses the above repeat rates to bound the entropy of a noise source.

Secondly, the statistics presented result in a loss of information about the noise source; however, there should be a computational (in terms of data required) advantage from this information loss. For example, one may have looser bounds on the entropy, but require less data to produce these bounds. The presented statistics leverage the birthday paradox to accommodate smaller data sets.

Thirdly, we desire efficient tests based on the statistics to be easy to implement. To address this criterion, we present tests that perform local measurements, requiring less memory and computation.

Finally, we desire that the collection of entropy estimates produced from the tests be robust with respect to a larger class of probability distributions (e.g. some non-*iid* sources). Assessment of the final criterion is beyond the scope of this paper.

We selected statistics based on the frequency of repeated outputs of the noise source, since they tend to satisfy the second the third criteria. In each of the next three subsections, we prove that the first criterion is satisfied by a different choice of statistic. The three statistics that we present are the **collision**, the **compression**, and the **partial collection**. The collision statistic measures the repeat rate without distinguishing different states: either the output is a repeat or it is not. The compression statistic is a weighted average of the repeat rates of every state. The partial collection statistic is a function of the average repeat rate of every state in a block of output without distinguishing the states present in a block. Thus the collision statistic is the most local measurement, the compression statistic is the most global, and the partial collection is a hybrid of the other two. We will make these descriptions precise in the subsequent subsections.

In the examples, we again assume *iid* output of entropy sources; we label the sequence

of output as $\{X_i\}_{i=1}^{t}$. In these examples, the entropy function that we are interested in is min-entropy, H_∞. The one-parameter family for the lower bound is the near-uniform family defined below.

Definition 4.1. We call the following one-parameter family of probability distributions parameterized by $\theta \in [0,1]$ on n states the **near-uniform** family:

$$\mathbf{p}_\theta[Z=i] = \begin{cases} \theta & i = i_1, \\ \frac{1-\theta}{n-1} & \text{otherwise.} \end{cases} \tag{4.1}$$

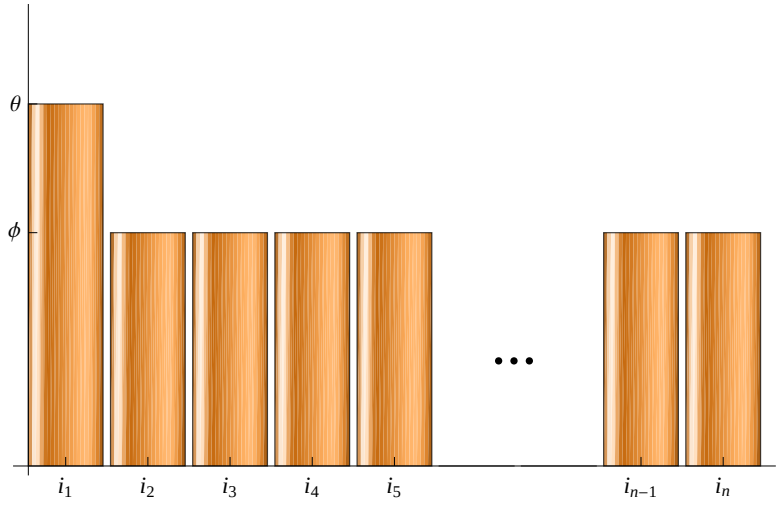

Figure 1: Probability mass function for the near-uniform family. Here $\varphi = \frac{1-\theta}{n-1}$.

In these examples, the one-parameter family for the upper bound is the inverted near-uniform family:

Definition 4.2. We call the following one parameter family of probability distributions parameterized by $\psi \in [0,1]$ on n states the **inverted near-uniform** family:

$$\mathbf{p}_\psi[Z=i] = \begin{cases} \psi & i \in \left\{i_1, \cdots, i_{\lfloor \frac{1}{\psi} \rfloor}\right\}, \\ 1 - \left\lfloor \frac{1}{\psi} \right\rfloor \psi & i = i_{\lfloor \frac{1}{\psi} \rfloor + 1}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

4.1. Entropic Example: Collision Statistic

The collision statistic computes the mean time to first collision of a sequence of output.

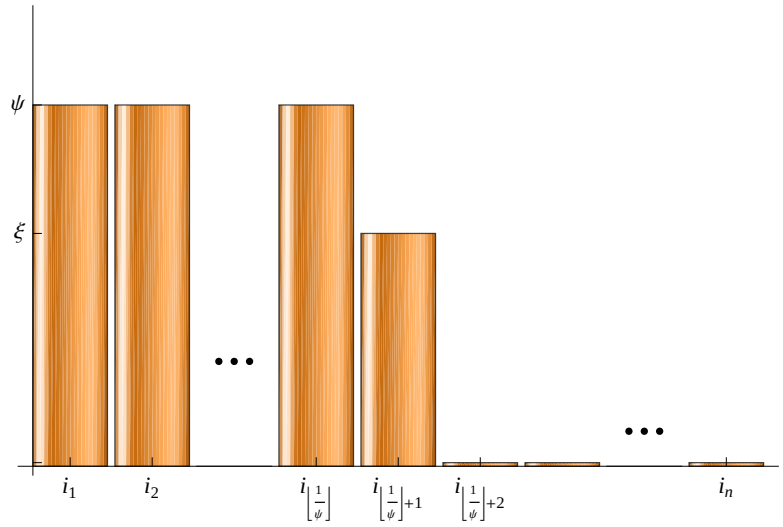

Figure 2: Probability mass function for the inverted near-uniform family. Here $\xi = 1 - \left\lfloor \frac{1}{\psi} \right\rfloor \psi$.

Statistic 4.3 (Collision)**.** *Generate a sequence* $\{X_s\}_{s=1}^{t}$ *of* iid *outputs from the entropy source. Let us define a sequence of collision times,* $\{T_i\}_{i=0}^{k}$*, as follows:* $T_0 = 0$ *and*

$$T_i = \min\{j > T_{i-1} : \textit{There exists } m \in (T_{i-1}, j) \textit{ such that } X_j = X_m\}. \tag{4.3}$$

The **collision statistic** *is the the average of differences of collision times,*

$$S_k = \frac{1}{k} \sum_{i=1}^{k} T_i - T_{i-1} = \frac{T_k}{k}. \tag{4.4}$$

Theorem 4.1. *The collision statistic is entropic with respect to* H_∞ *with the optimal distribution being a near-uniform distribution.*

We decompose the proof of the above theorem into lemmas stating monotonicity, convexity, and surjectivity.

Lemma 4.2 (Collision statistic monotonicity)**.** *Let* $\mathbf{p}_\theta$ *be the near-uniform family on* n *states and let* S_k *be the collision statistic. The expected value of the collision statistic,* $\mathbb{E}_{\mathbf{p}_\theta}(S_k)$*, is decreasing with respect to the parameter* θ *for* $\theta \in (1/n, 1]$ *:*

$$\frac{d}{d\theta} \mathbb{E}_{\mathbf{p}_\theta}[S_k] < 0 \textit{ for } \theta \in \left(\frac{1}{n}, 1 \right]. \tag{4.5}$$

Proof. Let $\mathbf{p}$ be written as a vector of probabilities $(p_1, \cdots, p_n)$. We explicitly compute the expected value of the collision statistic $\mathbb{E}_{\mathbf{p}}(S_k)$ as follows:

$$\mathbb{E}_{\mathbf{p}}[S_k] = \frac{1}{k} \mathbb{E}_{\mathbf{p}}[T_k], \tag{4.6}$$

$$= \mathbb{E}_{\mathbf{p}}[T_1], \tag{4.7}$$

by independence of the outputs. Since it is easy to compute the probability that a collision has not occurred at a specified time, we rewrite the expected value of the statistic as follows:

$$\mathbb{E}_{\mathbf{p}}[S_k] = \mathbb{E}_{\mathbf{p}}[T_1] = \sum_{i=1}^{n+1} \mathbb{P}[T_1 > i] = \sum_{i=1}^{n+1} P_i, \tag{4.8}$$

where

$$P_i = \mathbb{P}[\text{no collisions have occurred after } i \text{ outputs}] \tag{4.9}$$

Applying the above formula to the near-uniform family with $\varphi = \frac{1-\theta}{n-1}$, we have

$$P_i = i! \left(\binom{n-1}{i-1} \theta \varphi^{i-1} + \binom{n-1}{i} \varphi^i \right), \tag{4.10}$$

$$= i! \binom{n}{i} \left(\frac{i}{n} \theta \varphi^{i-1} + \left(1 - \frac{i}{n}\right) \varphi^i \right), \tag{4.11}$$

for $i \in \{0, 1, 2, \cdots, n\}$. Differentiating term by term, we have

$$\begin{aligned} \frac{d}{d\theta} P_i &= \frac{i! i \varphi^{i-2}}{n(n-1)} \binom{n}{i} \left((1 - i\theta) - \frac{n-i}{n-1}(1-\theta) \right), \\ &= \frac{i! i \varphi^{i-2}}{n(n-1)^2} \binom{n}{i} (i-1)(1 - n\theta). \end{aligned}$$

For each value of $i \in \{0, 1, 2, \cdots, n\}$, the term $\frac{d}{d\theta} P_i$ is negative for $\theta > \frac{1}{n}$. This completes the proof of Lemma 4.2. □

In many cases, we desire to efficiently compute $\mathbb{E}_{\mathbf{p}_\theta}[S]$. The following proposition reduces the computation to that of an incomplete gamma function, $\Gamma(a, z)$.

Proposition 4.4 (Computation)**.** *Using the above notation, we have*

$$\mathbb{E}_{\mathbf{p}_\theta}[S] = \theta \varphi^{-2} \left(1 + \frac{1}{n} \left(\theta^{-1} - \varphi^{-1} \right) \right) F(\varphi) - \theta \varphi^{-1} \frac{1}{n} \left(\theta^{-1} - \varphi^{-1} \right), \tag{4.12}$$

where

$$F(1/z) = \Gamma(n+1, z) z^{-n-1} e^z. \tag{4.13}$$

Proof. We simplify Equation 4.10 as follows:

$$\begin{aligned}
\sum_{i=0}^{n} P_i &= \sum_{i=0}^{n} i!\binom{n}{i}\left(\frac{i}{n}\theta\varphi^{i-1} + \left(1-\frac{i}{n}\right)\varphi^i\right), \\
&= \sum_{i=0}^{n} \frac{n!}{(n-i)!}\left(\frac{i\theta}{n}\varphi^{i-1} + \left(1-\frac{i}{n}\right)\varphi^i\right), \\
&= \sum_{s=0}^{n} \frac{n!}{s!}\left(\frac{(n-s)\theta}{n}\varphi^{n-s-1} + \left(1-\frac{n-s}{n}\right)\varphi^{n-s}\right), \\
&= n!\varphi^n \sum_{s=0}^{n} \frac{1}{s!}\left(\frac{(n-s)\theta}{\varphi n}\varphi^{-s} + \frac{s}{n}\varphi^{-s}\right), \\
&= A_0 \sum_{s=0}^{n} \frac{\varphi^{-s}}{s!} + A_1 \sum_{s=0}^{n} \frac{s\varphi^{1-s}}{s!}, \\
&= A_0 \sum_{s=0}^{n} \frac{\varphi^{-s}}{s!} + A_1 \sum_{r=0}^{n-1} \frac{\varphi^{-r}}{r!}, \\
&= (A_0 + A_1) \sum_{s=0}^{n} \frac{\varphi^{-s}}{s!} - \frac{\varphi^{-n}}{n!}A_1
\end{aligned}$$

where

$$\begin{aligned}
A_0 &= n!\varphi^{n-1}\theta, \\
A_1 &= n!\varphi^{n-1}\theta\left(\frac{\theta^{-1}-\varphi^{-1}}{n}\right).
\end{aligned}$$

Using the relation (see [3] §26.4)

$$\frac{\Gamma(n+1,\varphi^{-1})}{\Gamma(n+1)} = \sum_{s=0}^{n} e^{-1/\varphi}\frac{\varphi^{-s}}{s!}, \tag{4.14}$$

we obtain:

$$\sum_{s=0}^{n} \frac{\varphi^{-s}}{s!} = \frac{\varphi^{-n-1}}{n!}F(\varphi). \tag{4.15}$$

Then we have

$$\begin{aligned}
\sum_{i=0}^{n} P_i &= (A_0 + A_1)\frac{\varphi^{-n-1}}{n!}F(\varphi) - \frac{\varphi^{-n}}{n!}A_1, \\
&= \theta\varphi^{-2}\left(1 + \frac{1}{n}\left(\theta^{-1}-\varphi^{-1}\right)\right)F(\varphi) - \theta\varphi^{-1}\frac{1}{n}\left(\theta^{-1}-\varphi^{-1}\right),
\end{aligned}$$

as desired. □

There are many efficient algorithms for computing $F(\varphi)$ including the following continued fraction (see [3] §6.5.31) :

$$F(1/z) = \cfrac{1}{z + \cfrac{-n}{1 + \cfrac{1}{z + \cfrac{1-n}{1 + \cfrac{2}{z + \cfrac{2-n}{\cdots}}}}}}. \tag{4.16}$$

Lemma 4.3 (H_∞ monotonicity)**.** *Let $\mathbf{p}_\theta$ be the near-uniform family on n states. The value of min-entropy, H_∞, is decreasing with respect to the parameter $\theta \in (1/n, 1]$.*

Proof. The $\log_2$ function is monotonic. This completes the proof. □

Lemma 4.4. *Let T denote the first collision time of a sequence of* iid *outputs from a probability distribution $\mathbf{p}$ on a finite number of states. The value of $\mathbb{E}_{\mathbf{p}}[T]$ is maximized by the uniform distribution.*

Proof. Suppose there are n states in the probability space. Let $\mathbf{p}$ be written as a vector of probabilities $(p_1, \cdots, p_n)$. To show that the uniform distribution maximizes $\mathbb{E}_{\mathbf{p}}[T]$, it is sufficient to show that

$$\mathbb{E}_{\mathbf{p}'}[T] \geq \mathbb{E}_{\mathbf{p}}[T], \tag{4.17}$$

where

$$\mathbf{p}' = (\frac{1}{2}(p_1 + p_2), \frac{1}{2}(p_1 + p_2), p_3, \cdots, p_n). \tag{4.18}$$

Let $\mathbb{P}[T = t]$ denote the probability that the first collision occurs at time t; then we have,

$$\mathbb{P}[T = t] = \sum_{\mathbf{i} \in \mathcal{I}_t} \prod_{j=1}^{t} p_{i_j}, \tag{4.19}$$

where

$$\mathcal{I}_t = \left\{(i_1, \cdots, i_t) \in \{1, 2, \cdots, n\}^t | i_j \neq i_k \forall j \neq k\right\}. \tag{4.20}$$

Replacing p_1 and p_2 with their average will affect only the terms in the sum containing p_1, p_2, or both. By symmetry, any term containing p_1 or p_2 but not both has an analog that offsets the difference, so it suffices to consider what happens for $\mathbb{P}[T = t]$ when terms containing both p_1 and p_2 in the product are changed. We have

$$\left(\frac{p_1 + p_2}{2}\right)^2 - p_1 p_2 = \left(\frac{p_1 - p_2}{2}\right)^2 \geq 0. \tag{4.21}$$

Thus, $\mathbb{P}[T = t]$ increases. Since

$$\mathbb{E}[T] = \sum_t t\mathbb{P}[T = t], \tag{4.22}$$

we have $\mathbb{E}[T]$ also increases. □

Lemma 4.5 (Collision H_∞ convexity)**.** *Let $\mathcal{P}_H$ be the set of discrete probability distributions on n states with min-entropy H:*

$$\mathcal{P}_H = \{\mathbf{p} : H_\infty(\mathbf{p}) = H\}. \tag{4.23}$$

The near-uniform distribution, $\mathbf{p}_\theta$, maximizes the expected collision time over all $\mathbf{p} \in \mathcal{P}_H$.

Proof. Suppose we have a collection of probability distributions, $\mathcal{P}_H$, such that each probability distribution, $\mathbf{p} \in \mathcal{P}_H$, has min-entropy H. Without loss of generality, let p_n be the probability of the most likely state. Since the min-entropy is constant on the set $\mathcal{P}_H$, the value of p_n is also constant on the collection $\mathcal{P}_H$.

Let i_j and i_k represent two states distinct from state i_n. The proof of Lemma 4.4 also shows that replacing the probabilities of i_j and i_k with their average value increases the expected mean collision time. This completes the proof. □

Lemma 4.6 (Collision surjectivity)**.** *Let $\mathbf{p}$ be a probability distribution on n states. There exist values $\theta, \psi \in [\frac{1}{n}, 1]$ such that the near-uniform and inverted near-uniform probability distributions $\mathbf{p}_\theta$ and $\mathbf{p}_\psi$ satisfy the following:*

$$\begin{aligned} H_\infty(\mathbf{p}) &= H_\infty(\mathbf{p}_\theta), \\ \mathbb{E}_{\mathbf{p}}[S] &= \mathbb{E}_{\mathbf{p}_\psi}[S], \end{aligned}$$

where S is the collision statistic.

Proof. The functions H_∞ and $\mathbb{E}[S]$ are continuous with respect to the parameters θ and ψ. Invoking Lemma 4.4, we have $\mathbb{E}[S]$ is maximized when $\psi = \frac{1}{n}$. The choice of $\theta' = 1$ minimizes $\mathbb{E}[S]$. By the Intermediate Value Theorem, we have surjectivity of $\mathbb{E}[S]$. Surjectivity of H_∞ is trivial. □

Corollary 4.5. *Let S be the collision statistic. Then we have $-S$ is entropic with respect to $-H_\infty$ with the inverted near-uniform distribution being optimal.*

Proof. Surjectivity follows from the Lemma 4.6. Monotonicity follows from the chain rule and the calculations in Lemma 4.2:

$$\frac{d}{d\varphi}P_i = \frac{d}{d\theta}P_i\frac{d\theta}{d\varphi} = -(n-1)\frac{d}{d\theta}P_i. \tag{4.24}$$

Convexity follows from the fact that the extrema of the P_i occur on the boundary. □

We combine the above results in a theorem.

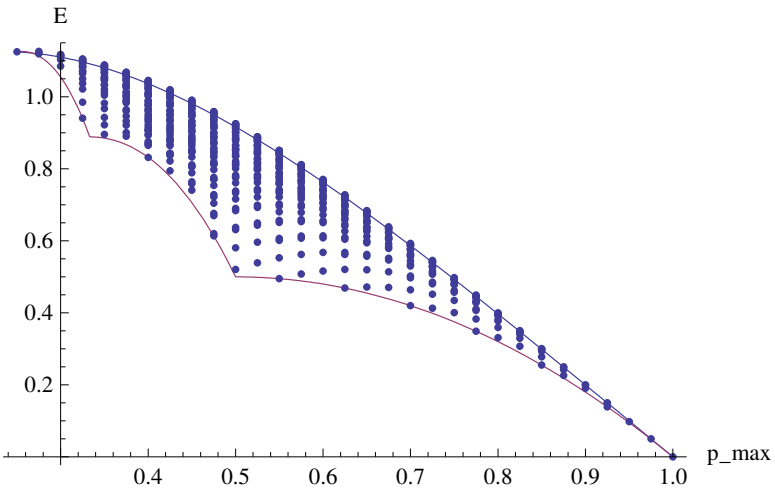

Figure 3: Example of entropy bounds for collision test on four states. The upper bound of $\mathbb{E}_{\mathbf{p}}(S)$ is from the near-uniform distribution. The lower bound of $\mathbb{E}_{\mathbf{p}}(S)$ is from the inverted near-uniform distribution. Sample distributions are also plotted.

Theorem 4.7. *Let* $\mathbf{p}$ *be a probability distribution on* n *states and let* S *be the collision statistic. Let us choose* $\theta, \psi \in [\frac{1}{n}, 1]$ *such that* $\mathbf{p}_\theta$ *is a near-uniform distribution with*

$$\mathbb{E}_{\mathbf{p}_\theta}[S] = \mathbb{E}_{\mathbf{p}}[S], \tag{4.25}$$

and $\mathbf{p}_\psi$ *is an inverted near-uniform distribution with*

$$\mathbb{E}_{\mathbf{p}_\psi}[S] = \mathbb{E}_{\mathbf{p}}[S]. \tag{4.26}$$

Then we can bound the min-entropy of $\mathbf{p}$ *above and below:*

$$H_\infty(\mathbf{p}_\theta) \leq H_\infty(\mathbf{p}) \leq H_\infty(\mathbf{p}_\psi). \tag{4.27}$$

4.2. Entropic Example: Compression Statistic

We analyze the Universal Maurer Statistic ([5]) in the entropic setting. This compression test has two phases: a dictionary creation phase and a computation phase. We show that the entropic nature applies to both a finite creation and an infinite creation. We now define the compression statistic of entropy.

Statistic 4.6 (Compression (Maurer))**.** *Generate a sequence* $\{X_s\}_{s=1}^t$ *of outputs from the entropy source. Partition the outputs into two disjoint groups: the dictionary group comprises the first* r *outputs and the test group comprises the next* k *outputs. The* **compression statistic**, $S_{r,k}$, *is the average of the following compression values computed over the test group:*

$$S_{r,k} = \frac{1}{k} \sum_{s=r+1}^{r+k} \log_2 A_s(X_s), \tag{4.28}$$

where

$$A_s = \begin{cases} s & \text{if } X_s \neq X_u \forall 0 < u < s, \\ s - \max\{u < s : X_s = X_u\} & \text{otherwise.} \end{cases} \tag{4.29}$$

Theorem 4.8. *The compression statistic is entropic with respect to H_∞ with the near-uniform distribution being the optimal distribution.*

Proof. We will substitute the function $\zeta(x)$ for $\log_2(x)$ for generality. When computing the expected value of the compression statistic, $\mathbb{E}_{\mathbf{p}_\theta}[S_{r,k}]$, we decompose the computation into times that a collision had been detected:

$$\begin{aligned} \mathbb{E}_{\mathbf{p}}[S_{r,k}] &= \frac{1}{k}\sum_{s=r+1}^{r+k} \mathbb{E}[\zeta(A_s(X_s))], && (4.30) \\ &= \frac{1}{k}\sum_{s=r+1}^{r+k}\left(\sum_{u=1}^{s}\zeta(u)\mathbb{P}[A_s = u]\right), && (4.31) \\ &= \frac{1}{k}\sum_{s=r+1}^{r+k}\left(\sum_{u=1}^{s}\zeta(u)\sum_i \mathbb{P}[A_s = u \cap X_s = i]\right), && (4.32) \\ &= \frac{1}{k}\sum_i\left(\sum_{s=r+1}^{r+k}\sum_{u=1}^{s}\zeta(u)\mathbb{P}[A_s = u \cap X_s = i]\right), && (4.33) \\ &= \sum_i G(p_i), && (4.34) \end{aligned}$$

where

$$G(p_i) = \frac{1}{k}\sum_{s=r+1}^{r+k}\sum_{u=1}^{s}\zeta(u)\mathbb{P}[A_s = u \cap X_s = i], \tag{4.35}$$

and

$$\mathbb{P}[A_s = u \cap X_s = i] = \begin{cases} p_i^2(1-p_i)^{u-1} \text{ if } u < s, \\ p_i(1-p_i)^{s-1} \text{ if } u = s. \end{cases} \tag{4.36}$$

Applying the above formula to the near-uniform distribution on n states we have

$$\mathbb{E}_{\mathbf{p}_\theta}[S_{r,k}] = G(\theta) + (n-1)G(\varphi), \tag{4.37}$$

where $\varphi = \frac{1-\theta}{n-1}$.

Monotonicity: To show monotonicity, we desire to show

$$\frac{d}{d\theta}\mathbb{E}_{\mathbf{p}_\theta}[S_{r,k}] < 0, \tag{4.38}$$

for $\theta \in \left(\frac{1}{n}, 1\right)$. Differentiating term by term, we have

$$\frac{d}{d\theta}\mathbb{E}_{\mathbf{p}_\theta}[S] = G'(\theta) + (n-1)G'(\varphi)\frac{d\varphi}{d\theta}, \tag{4.39}$$

$$= G'(\theta) - G'(\varphi). \tag{4.40}$$

Treating G as a function on $[0,1]$, we can write $G(x)$ in powers of $(1-x)$:

$$G(x) = \frac{1}{k}\sum_{s=r+1}^{r+k}\left(\sum_{u=0}^{s+1} g_u(1-x)^u\right), \tag{4.41}$$

where

$$g_u = \begin{cases} \zeta(1) & \text{if } u = 0, \\ \zeta(2) - 2\zeta(1) & \text{if } u = 1, \\ \zeta(u+1) - 2\zeta(u) + \zeta(u-1) & \text{if } u \in \{2, \cdots, s-1\}, \\ \zeta(u-1) - \zeta(u) & \text{if } u = s. \end{cases} \tag{4.42}$$

Differentiating term by term, we have

$$G''(x) = \frac{1}{k}\sum_{s=r+1}^{r+k}\left(\sum_{u=2}^{s+1} u(u-1)g_u(1-x)^{u-2}\right). \tag{4.43}$$

To show $G''(x) < 0$, we can impose some restrictions on ζ so that the terms in the expansion of $G''(x)$ are negative. If ζ is concave down and increasing (as $\log_2$ is), then $g_u < 0$ for $u = 2, \cdots, s$. Thus we have,

$$G''(x) < 0, \tag{4.44}$$

for $x \in (0,1)$ and

$$G'(\theta) - G'(\varphi) < 0, \tag{4.45}$$

as desired. Since $\zeta = \log_2$ is concave down, increasing, and positive, we have completed the proof of monotonicity.

Convexity: Convexity follows directly from Equation 4.44.

Surjectivity: Surjectivity follows from convexity and the extremal nature of the near-uniform distribution.

Applying the Entropy Theorem (Theorem 3.1), we have the desired result. This completes the proof. □

Corollary 4.7. *Suppose we replace the* $\log_2$ *function in the compression statistic with the function* ζ *to define a new statistic* S*. If* ζ *is concave down, increasing, and positive then we have the statistic* S *is entropic with respect to* H_∞ *with the near-uniform distribution being the optimal distribution.*

Corollary 4.8. *Suppose we replace the* $\log_2$ *function in the compression statistic with the function* ζ *to define a new statistic* $S_{r,k}$*. If* ζ *is such that*

$$\frac{1}{k}\sum_{s=r+1}^{r+k}\left(\sum_{u=2}^{s+1}u(u-1)g_u(1-x)^{u-2}\right)<0, \tag{4.46}$$

for $x\in[0,1]$ *and where* g_u *is defined in terms of* ζ *in Equation 4.42, then we have the statistic* $S_{r,k}$ *is entropic with respect to* H_∞ *with the near-uniform distribution being the optimal distribution.*

Corollary 4.9. *Suppose one computes the dictionary over a sliding window of length* τ*:*

$$A_s=\begin{cases}\tau & \text{if } X_s\neq X_u\forall s-\tau<u<s,\\ \tau-\max\{u:s-\tau<u<s,X_s=X_u\} & \text{otherwise.}\end{cases} \tag{4.47}$$

The statistic using the new definition of A_s *is entropic with respect to* H_∞*.*

Corollary 4.10. *The negative of the compression statistic is entropic with respect to* $-H_\infty$*, with the inverted near-uniform distribution being the optimal distribution.*

We combine the above results in a theorem.

Theorem 4.9. *Let* $\mathbf{p}$ *be a probability distribution on* n *states and let* $S_{r,k}$ *be the compression statistic. Let us choose* θ *and* ψ *such that* $\mathbf{p}_\theta$ *is a near-uniform distribution with*

$$\mathbb{E}_{\mathbf{p}_\theta}[S_{r,k}]=\mathbb{E}_{\mathbf{p}}[S_{r,k}], \tag{4.48}$$

and $\mathbf{p}_\psi$ *is an inverted near-uniform distribution with*

$$\mathbb{E}_{\mathbf{p}_\psi}[S_{r,k}]=\mathbb{E}_{\mathbf{p}}[S_{r,k}]. \tag{4.49}$$

Then we can bound the min-entropy of $\mathbf{p}$ *above and below:*

$$H_\infty(\mathbf{p}_\theta)\leq H_\infty(\mathbf{p})\leq H_\infty(\mathbf{p}_\psi). \tag{4.50}$$

4.3. Entropic Example: Partial Collection Statistic

The partial collection statistic computes the expected number of distinct values observed in a fixed number of outputs.

Statistic 4.11 (Partial Collection)**.** *Generate a sequence* $\{X_s\}_{s=1}^t$ *of* iid *outputs from the entropy source. Let us partition the output sequence into non-overlapping sets of length* $k\in\mathbb{N}$*. Let* A_i *denote the number of distinct values in the* i^{th} *set. The* **partial collection statistic***,* S*, is the average of the number of distinct outputs,* A_i*:*

$$S=\frac{1}{\left\lfloor\frac{t}{k}\right\rfloor}\sum_{i=1}^{\left\lfloor\frac{t}{k}\right\rfloor}A_i. \tag{4.51}$$

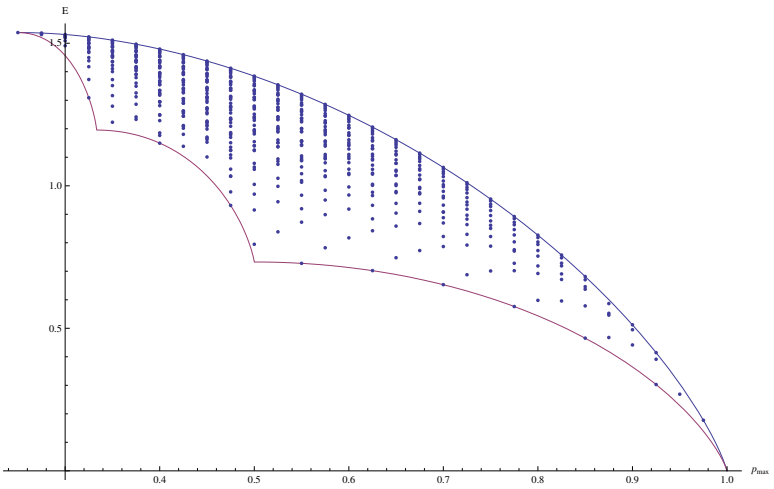

Figure 4: Example of entropy bounds for compression test on four states. The upper bound of $\mathbb{E}_{\mathbf{p}}(S)$ is from the near-uniform distribution. The lower bound of $\mathbb{E}_{\mathbf{p}}(S)$ is from the inverted near-uniform distribution. Sample distributions are also plotted.

Theorem 4.10. *Let* $\mathbf{p}$ *be a probability distribution on* n *states and let* S *be the partial collection statistic. Then there exist values of* θ *and* ψ *such that* $\mathbf{p}_\theta$ *is a near-uniform distribution with*

$$\mathbb{E}_{\mathbf{p}_\theta}[S] = \mathbb{E}_{\mathbf{p}}[S], \tag{4.52}$$

and $\mathbf{p}_\psi$ *is an inverted near-uniform distribution with*

$$\mathbb{E}_{\mathbf{p}_\psi}[S] = \mathbb{E}_{\mathbf{p}}[S]. \tag{4.53}$$

Furthermore, we can bound the min-entropy of $\mathbf{p}$ *above and below:*

$$H_\infty(\mathbf{p}_\theta) \leq H_\infty(\mathbf{p}) \leq H_\infty(\mathbf{p}_\psi). \tag{4.54}$$

Proof. We first compute the expected value of the partial collection statistic by computing the probability that the j^{th} state is in the partial collection and summing:

$$\mathbb{E}_{\mathbf{p}}[S] = \sum_j 1 - (1-p_j)^k, \tag{4.55}$$

where k is the length of each non-overlapping set. For the optimal distributions we have

$$\mathbb{E}_{\mathbf{p}_\theta}[S] = 1 - (1-\theta)^k + (n-1)\left(1 - (1-\varphi)^k\right), \tag{4.56}$$

$$\mathbb{E}_{\mathbf{p}_\psi}[S] = \left\lfloor \frac{1}{\psi} \right\rfloor \left(1 - (1-\psi)^k\right) + 1 - (1-\xi)^k, \tag{4.57}$$

where $\varphi = \frac{1-\theta}{n-1}$ and $\xi = 1 - \left\lfloor \frac{1}{\psi} \right\rfloor \psi$. Monotonicity, convexity, and surjectivity follow from direct computation. □

4.4. Entropic Example: Rényi Entropy

In this section, we apply the above technique to bound min-entropy based on a measure of a Rényi entropy. We present a theorem that strengthens a classical result of ordering Rényi entropies for a specified probability distribution (i.e. $H_\infty \leq H_\alpha$ if $\alpha \geq 1$) by giving a formula for the optimizing distribution. This construction yields explicit bounds on the min-entropy given other Rényi entropy measurements.

Theorem 4.11. *Let* $\mathbf{p}$ *be a probability distribution on* n *states and let* $\alpha \in [1, \infty]$. *Choose* $\theta, \psi \in [\frac{1}{n}, 1]$ *such that the near-uniform and inverted near-uniform distributions,* p_θ, p_ψ, *satisfy the following:*

$$H_\alpha(\mathbf{p}_\theta) = H_\alpha(\mathbf{p}_\psi) = H_\alpha(\mathbf{p}). \tag{4.58}$$

Then we have,

$$H_\infty(\mathbf{p}_\theta) \leq H_\infty(\mathbf{p}) \leq H_\infty(\mathbf{p}_\psi). \tag{4.59}$$

Proof. We model the proof on the proof of Theorem 3.1, replacing expected value of the statistic with the value of the H_α. One may rigorously obtain the same result as a limiting process of expected values of a family of statistics; however, the proof of Theorem 3.1 generalizes to function evaluations.

Previously, we have shown monotonicity and surjectivity of H_∞. We now desire to show monotonicity of H_α. Let $\mathbf{p}_\theta$ be a near-uniform probability distribution; then we have for $\alpha \neq 1$ and $\alpha \neq \infty$,

$$H_\alpha(\mathbf{p}_\theta) = \frac{1}{1-\alpha} \log_2 \left(\theta^\alpha + (n-1) \left(\frac{1-\theta}{n-1} \right)^\alpha \right). \tag{4.60}$$

Differentiating with respect to θ, we have

$$\frac{d}{d\theta} H_\alpha(\mathbf{p}_\theta) = \frac{\alpha \left(\theta^{\alpha-1} - \varphi^{\alpha-1} \right)}{(1-\alpha) \ln 2 \left(\theta^\alpha + (n-1)\varphi^\alpha \right)}, \tag{4.61}$$

where $\varphi = \frac{1-\theta}{n-1}$. For $\theta \in \left(\frac{1}{n}, 1\right]$, the numerator is positive and the denominator is negative, proving monotonicity of $H_\alpha(\mathbf{p}_\theta)$. In the limiting cases ($\alpha \in \{1, \infty\}$), the result also holds:

$$\begin{aligned} \frac{d}{d\theta} H_1(\mathbf{p}_\theta) &= -\frac{d}{d\theta} \left(\theta \log_2 \theta + (n-1)\varphi \log_2 \varphi \right), \\ &= \frac{-1}{\ln 2} \left(\ln \theta + 1 - \ln \varphi - 1 \right), \\ &= -\log_2 \left(\frac{\theta}{\varphi} \right). \end{aligned}$$

Now we desire to show convexity. Assume that $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ where $p_k \leq p_n$ for $k = 1, \cdots, n$. Replacing two probabilities with their average, we obtain a new probability distribution $\mathbf{p}' = (\frac{p_1+p_2}{2}, \frac{p_1+p_2}{2}, p_3, \cdots, p_n)$. It is sufficent to show

$$H_\alpha(\mathbf{p}) \leq H_\alpha(\mathbf{p}'). \tag{4.62}$$

Taking the difference, we have

$$\frac{1}{1-\alpha}\left(\log_2\left(p_1^\alpha + p_2^\alpha + \sum_{k=3}^{n} p_k^n\right) - \log_2\left(2\left(\frac{p_1+p_2}{2}\right)^\alpha + \sum_{k=3}^{n} p_k^n\right)\right). \tag{4.63}$$

Since the $\log_2$ is monotonic, it suffices to show

$$p_1^\alpha + p_2^\alpha - 2\left(\frac{p_1+p_2}{2}\right)^\alpha \leq 0, \tag{4.64}$$

with equality iff $p_1 = p_2$, which holds by convexity of $f(x) = x^\alpha$ for $\alpha > 1$. Explicitly computing the limiting case, we have

$$H_1(\mathbf{p}) - H_1(\mathbf{p}') = p_1 \log_2 p_1 + p_2 \log_2 p_2 - (p_1+p_2)\log_2\left(\frac{p_1+p_2}{2}\right). \tag{4.65}$$

By convexity of $f(x) = x\log_2 x$, we have

$$H_1(\mathbf{p}) - H_1(\mathbf{p}') \leq 0, \tag{4.66}$$

with equality iff $p_1 = p_2$.

Surjectivity of H_α follows from the Rényi entropies' bounds of 0 and $\log_2 n$. The proof of Theorem 3.1 yields the desired lower bound when one replaces $\mathbb{E}_x[S]$ with $H_\alpha(x)$. The upper bound follows for a similar computation applied to $-H_\alpha$ and $\mathbf{p}_\psi$. □

5. Comparing Entropy Statistics

The power of entropically bounding statistics is that they are quantitatively comparable. The dual-sided entropic property of an entropically bounding statistic allows us to define $H_{\max}$ as a function of $H_{\min}$.

Definition 5.1. Suppose S entropically bounds the entropy function H. Let m be a measurement of the statistic S and define $H_{\max}(m)$ and $H_{\min}(m)$ as follows:

$$H_{\max}(m) = \max_{\mathbf{p}\in\mathcal{P}_S(m)} \{H(\mathbf{p})\}, \tag{5.1}$$

$$H_{\min}(m) = \min_{\mathbf{p}\in\mathcal{P}_S(m)} \{H(\mathbf{p})\}. \tag{5.2}$$

We define the **spread of** S **at** m as the function:

$$\text{spread}(S,m) = H_{\max}(m) - H_{\min}(m). \tag{5.3}$$

Since $H_{\min}$ is monotonic, its inverse, $H_{\min}^{-1}$, is well defined. We define the **spread of** S as

$$\text{spread}_S(h) = \text{spread}\left(S, H_{\min}^{-1}(m)\right). \tag{5.4}$$

Definition 5.2. The **average spread of** S is the average of the spread over all possible values of entropy:

$$\overline{\text{spread}}_S = \frac{\int \text{spread}_S(h)\ dh}{\int\ dh}, \tag{5.5}$$

where the limits of integration are defined by the surjectivity of the entropic statistic.

Definition 5.3. Let $\varepsilon > 0$. Let $B_\varepsilon(m)$ denote the closed ball around m of radius ε in a metric space and

$$H_{\max,\varepsilon}(m) = \max\left\{H(\mathbf{p})\ :\ \mathbb{E}_{\mathbf{p}}\left[S\right] \in B_\varepsilon(m)\right\}, \tag{5.6}$$

$$H_{\min,\varepsilon}(m) = \min\left\{H(\mathbf{p})\ :\ \mathbb{E}_{\mathbf{p}}\left[S\right] \in B_\varepsilon(m)\right\}, \tag{5.7}$$

Then the ε**-dilation spread of** S **at** m is defined as

$$\text{dspread}_\varepsilon(S, m) = H_{\max,\varepsilon}(m) - H_{\min,\varepsilon}(m). \tag{5.8}$$

If $H_{\min,\varepsilon}$ is well defined monotonic, we define the ε**-dilation spread of** S and the **average** ε**-dilation spread of** S as follows:

$$\text{dspread}_{\varepsilon,S}(h) = \text{dspread}_\varepsilon(S, (H^{-1}_{\min,\varepsilon}(m)), \tag{5.9}$$

$$\overline{\text{dspread}}_{\varepsilon,S} = \frac{\int \text{dspread}_{\varepsilon,S}(h)\ dh}{\int dh}. \tag{5.10}$$

The spread of a statistic is useful when comparing two statistical tests. Moreover, one could hope to compute the best statistical test to perform on a source to estimate the entropy of the output. From our analysis, a statistical test whose expected value is a monotone function of the entropy yields an identically zero spread. However, this optimum neglects the variance inherent in a finite collection of data.

Problem 5.4. *Let $\mathcal{S}$ be a collection of statistics that each entropically bounds the entropy function H. Given an $\varepsilon > 0$, suppose the average ε-dilation spread is well defined for all $S \in \mathcal{S}$. Find the statistic S_ε that minimizes the average ε-dilation spread.*

Normalization remains an issue in solving the above problem. In the limit as ε tends to zero, we see that any statistic with the expected value being a monotonic function of entropy becomes optimal. However, other statistics might be more robust with respect to ε. We leave this topic to future research.

In particular, the estimation of H_∞ directly from a frequency count is optimal when ε is zero. However, for $\varepsilon > 0$, other statistics may have a smaller spread.

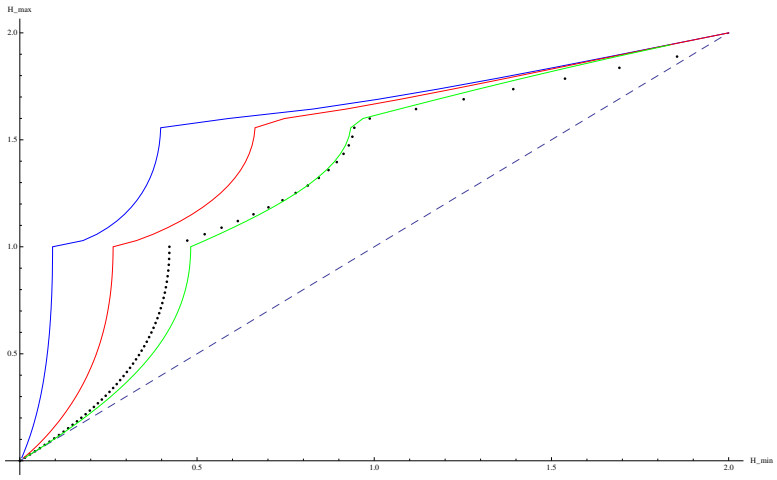

Figure 5: For the compression test on four states, $H_{\max}$ is plotted as a function of $H_{\min}$ for the following choices of ζ: blue is for $\zeta(x) = \sqrt{x}$, red is for $\zeta(x) = \log_2 x$, green is for $\zeta(x) = 1 - e^{-x}$. The dashed purple line is the identity bound. Additionally, the same function is plotted for the collision test as black points.

5.1. Examples: Compression Spread versus Collision Spread and Optimal φ

Even in the case where ε is zero, Problem 5.4 introduces a way to compare seemingly uncomparable statistics. We can tighten the comparison using a pointwise comparison of the spread function. In Figures 5 and 6, we present a summary of the analysis of spread for the collision test and the compression test with three choices of the weighting function ζ.

Recall from the analysis of the compression statistic, there was a large amount of freedom for the weighting function: ζ had to be increasing and concave down. The default choice of ζ was the logarithm function $\log_2$. In this subsection, we compare the spread of several choices of ζ.

The data suggest that the collision statistic is a tighter statistic than the default compression statistic ($\zeta = \log_2$). However, we have only performed the analysis for a small number of states. Furthermore, it appears that by optimally choosing ζ, one can improve the compression statistic so that it is tighter than the collision statistic.

Problem 5.5. *Find the ζ that minimizes the average spread for the compression statistic such that $\zeta''(t) < 0$ and $\zeta'(t) > 0$.*

6. Entropy Measurements

Suppose we have an unknown probability distribution, $\mathbf{p}$, that produces a sequence of outputs. A statistic computed on the sequence of output will usually not be the expected value of the statistic. To accommodate this noise, we relax the constraint, generate a set of admissible distributions based on distance to the measured statistic, and solve the following

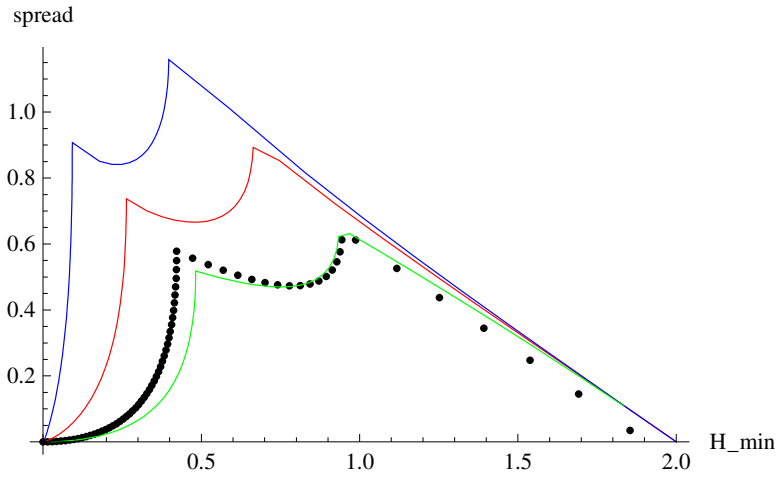

Figure 6: For the compression test on four states, the spread is plotted as a function of $H_{\min}$ for the following choices of ζ: blue is for $\zeta(x) = \sqrt{x}$, red is for $\zeta(x) = \log_2 x$, green is for $\zeta(x) = 1 - e^{-x}$. Additonally, the same function is plotted for the collision test as black points.

problem.

Problem 6.1. *Solve the constrained optimization problem:*

$$H_\varepsilon(m) = \min_{\mathbf{p} \in \mathcal{P}_{S,\varepsilon}(m)} H(\mathbf{p}), \tag{6.1}$$

where

$$\mathcal{P}_{S,\varepsilon}(m) = \{\mathbf{p} : \mathbb{E}_{\mathbf{p}}[S] \in B_\varepsilon(m)\}. \tag{6.2}$$

We present three types of entropy tests to solve Problem 6.1: entropic, dilation, and dilation with confidence. Based on the motivating pass/fail tests of full entropy, we continue to use the term "test."

6.1. Entropic Tests

If the statistic is entropic, the relaxed minimization problem effectively reduces to Problem 3.1 with a shifted value of m. The following entropic test shifts the value of m based on the number of samples of output.

Test 6.2 (Entropic)**.** *Generate a sequence of output from the entropy source. Let $S_1, S_2, \cdots, S_k$ be a sequence of independent measurements of an entropic statistic S, with respect to the function H, computed from the sequence, with sample mean $\bar{S}$ and sample variance $\tilde{\sigma}^2$ over the k samples. Given a confidence level α, let us define an adjusted mean $\bar{S}'$ as follows:*

$$\bar{S}' = \bar{S} - \Phi^{-1}(1-\alpha)\frac{\tilde{\sigma}}{\sqrt{k}}. \tag{6.3}$$

where Φ is the standard normal cumulative distribution function. Define the entropy estimate H_{est} by minimizing the function H over all distributions on the same state space that have the expected value of the statistic S greater than or equal to $\bar{S}'$:

$$H_{est} = \min \left\{ H(\mathbf{p}) : \mathbb{E}_{\mathbf{p}}[S] \geq \bar{S}' \right\}. \tag{6.4}$$

6.2. Dilation Tests

The following is the general dilation test.

Test 6.3 (Dilation)**.** *Generate a sequence of output from the entropy source. Let H be an entropy function. Let $S_1, S_2, \cdots, S_k$ be a sequence of independent meaurements of a statistic, S, computed from the sequence, with sample mean $\bar{S}$. Given a tolerance level ε, define the entropy estimate H_{est} by minimizing the function H over the set of distributions $\mathcal{P}_{S,\varepsilon}(\bar{S})$:*

$$H_{est} = \min_{\mathbf{p} \in \mathcal{P}_{S,\varepsilon}(\bar{S})} H(\mathbf{p}). \tag{6.5}$$

The dilation test is easily adapted to vector-valued statistics. Since we have relaxed the monotonicity condition, the optimization problem may be intractable. We have not prescribed a method for choosing the value of ε. We can realize the entropic test as a special case of a dilation test with ε depending on the standard deviation of the statistic and the number of samples.

A standard result for Bernoulli(p) distributions motivates a further generalization.

Test 6.4 (Dilation with confidence)**.** *Generate a sequence of output from the entropy source. Let H be an entropy function. Let $S_1, S_2, \cdots, S_k$ be a sequence of measurements of a statistic, S, computed from the output sequence, with sample mean $\bar{S}$. Given a confidence level α and tolerance level ε, define a dilation collection of probability distributions, $\mathcal{P}_{\alpha,\varepsilon,S}(\bar{S})$ as follows,*

$$\mathcal{P}_{S,\alpha,\varepsilon}(\bar{S}) = \left\{ \mathbb{P} : \mathbb{P}\left[S \in B_\varepsilon\left(\bar{S}\right) \right] \geq \alpha \right\}, \tag{6.6}$$

where $\mathbb{P}$ is the probability measure for the computed statistic S (considered a random variable). Define the entropy estimate H_{est} by minimizing the function H over the set of distributions $\mathcal{P}_{S,\alpha,\varepsilon}(\bar{S})$:

$$H_{est} = \min_{\mathbb{P} \in \mathcal{P}_{S,\alpha,\varepsilon}(\bar{S})} H(\mathbb{P}). \tag{6.7}$$

In the dilation with confidence test, the choice of ε and of α control the set of admissible probability distributions over which to optimize. As the value of ε decreases, admissibility becomes more restrictive, resulting in an expected increase in the value of H_{est}.

Hoeffding's Inequality (see [4]) relates the value of the confidence level α to the tolerance level ε for Bernoulli(p) streams.

Theorem 6.1 (Hoeffding's Inequality)**.** *Let $X_1, \cdots, X_n \sim$ Bernoulli(p). Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\left[p \in B_\varepsilon(\bar{X}) \right] \geq \alpha, \tag{6.8}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\alpha = 1 - 2e^{-2n\varepsilon^2}$.

6.3. Multiple Tests

Often one computes several statistics on a data source and estimates the entropy based on all of the results. If one performs a test on each statistic, the estimate of the entropy will most likely be different for each test. In this section, we present ways to resolve the issue.

If all of the tests computed are of the same type (entropic, dilation, or dilation with confidence) using the same entropy function, then there are two simple approaches. The first simple approach is to be conservative: choose the minimum estimate as the estimate. The second simple approach is to choose the maximal estimate as the estimate. The maximal approach can be made rigorous when all of the tests are entropic using interval arithmetic on the concept of spread.

During each statistical test, one optimizes over a set of probability distributions satisfying constraints determined by the statistic. A general approach is to optimize over the intersection of the sets. In some cases, this optimization problem is similar to the individual optimizations. If all of the tests are entropic, then the surjectivity condition implies that the intersection is nonempty. For each dilation test, one can increase the tolerance level until the intersection is nonempty.

Alternatively, one could vectorize the statistic and apply a vectorized dilation test (rewriting entropic tests as dilation test). The difficult result here is that the optimization problem may be intractable.

A. Example Dilation Test: Frequency Statistic

The frequency statistic computes the entropy by modeling a probability distribution based on output of a random source.

Statistic A.1 (Frequency: min-entropy)**.** *Generate a sequence* $\{X_s\}_{s=1}^{t}$ *of outputs from the entropy source. For each state* i *of the output space estimate the probability of occupancy of state* i*:*

$$P_i = \frac{1}{t}\sum_{s=1}^{t} \chi_i(X_s), \tag{A.1}$$

where

$$\chi_i(X_s) = \begin{cases} 1 & \text{if } i = X_s, \\ 0 & \text{if } i \neq X_s. \end{cases} \tag{A.2}$$

The **min-entropy frequency statistic**, S*, is the negative logarithm of estimated probability of the most frequent state:*

$$S = -\log_2(\max_j P_j). \tag{A.3}$$

Even though the min-entropy frequency statistic is entropic by Theorem 4.11, data requirements suggest that one use a dilation with confidence test for the frequency statistic. In the case of *iid* output, we can realize the frequency statistic as a Bernoulli(p) process in each state and invoke Hoeffding's Inequality.

B. Collection Statistic

The collection statistic is a complement to the collision test. The statistic yields low entropy estimates for output streams that diversify quickly and yields high entropy estimates for output streams that fail to diversify quickly enough. The collection test is a sanity check that may increase the accuracy of entropy estimates of counters, linear feedback shift registers, and linear congruential generators; other statistics may overestimate the entropy from these type of sources.

Statistic B.1 (Collection)**.** *Generate a sequence* $\{X_s\}_{s=1}^{t}$ *of output from the entropy source. Let us define a sequence of collection times,* $\{T_i\}_{i=0}^{k}$*, as follows:* $T_0 = 0$ *and*

(B.1)

$$T_i = \min\left\{\{T_{i-1}+\Delta\} \cup \{j > T_{i-1} : \textit{every output state has occured in } \{X_m\}_{m=T_i+1}^{j}\}\right\},$$

where Δ *is the maximum allotted collection time. The* **collection statistic***, S, is the average of differences of collection times,*

$$S = \frac{1}{k}\sum_{i=1}^{k} T_i - T_{i-1} = \frac{T_k}{k}. \tag{B.2}$$

A major concern with the collection statistic is that it may require too much data to collect a full collection, let alone a statistically significant number of full collections. As min-entropy tends to zero, the expected time of completion tends to infinity. Most implementations bound the collection time, making it impossible to relate the collection statistic to min-entropy. In contrast, the partial collection test entropically bounds min-entropy.

C. Markovity and Dependencies

Independence is a tricky issue when it comes to entropy testing. Almost all of the rigorous results assume *iid* output of the source. One can construct entropy sources with dependencies in time and/or state such that the entropy tests overestimate the entropy instead of underestimating it. However, a large, diverse battery of tests minimize the probability that such a source's entropy is greatly overestimated. The tests provide a sanity check for the entropy estimate instead of a rigorous bound!

One difficulty in modeling sources with dependencies is the large data requirement to resolve the dependencies from sampling. The canonical example of dependent data is a Markov process: the output state only depends on the current state. We can use the Markov model as a template on which to project more complicated sources with dependencies.

The key component of estimating the entropy of a Markov process is the ability to accurately estimate the matrix of transition probabilities of the Markov process. When computing the min-entropy we can use a dilation with confidence test to minimize the required data. The technique sacrifices large data requirements for overestimations of unlikely transitions.

C.1. Dynamic Programming

Let $X(t)$ be a Markov process with transition matrix T. By estimating the probability of the initial distribution, one computes the largest probability of a fixed number of outputs, N, of the process and estimate the min-entropy of the outputs collectively. Explicitly, we minimize over all chains of N states labeled $i_0, \cdots, i_{N-1}$:

$$H_\infty(T, \mathbf{p}, n) = \min_{i_0,\cdots,i_{N-1}} -\log_2 \mathbb{P}[X_0 = i_0 \cap \cdots \cap X_{N-1} = i_{N-1}], \tag{C.1}$$

$$= \min_{i_0,\cdots,i_{N-1}} -\log_2 \left(p_{i_0} \prod_{j=0}^{N-1} T_{i_j,i_{j+1}} \right), \tag{C.2}$$

where p_{i_0} is the initial probability of outputing state i_0. We solve the optimization problem with standard dynamic programming techniques.

C.2. Bounds on the Transition Matrix

When one estimates the transition matrix from a fixed data set, the size of the data set significantly impacts the accuracy of the estimate. In particular, low probability transitions may not occur often in the data set. If one is able to overestimate the transition probability, we obtain an underestimate of min-entropy.

Proposition C.1. *Let $X(t)$ be a Markov process on n states with one-step transition matrix T, and initial distribution $\mathbf{p}$ where $H_\infty(T, \mathbf{p}, N)$ denotes the solution to the above dynamic programming problem. If $S \in [0,1]^{n\times n}$ is a matrix such that $S_{ij} \geq T_{i,j}$ for $i, j = 1, \cdots, n$, then we have*

$$H_\infty(S, \mathbf{p}, N) \leq H_\infty(T, \mathbf{p}, N). \tag{C.3}$$

Proof. In Equation C.1, we have

$$\prod_{j=0}^{N-1} T_{i_j,i_{j+1}} \leq \prod_{j=0}^{N-1} S_{i_j,i_{j+1}}. \tag{C.4}$$

The desired inequality follows directly from the decreasing property of the function $-\log_2$. □

It is interesting to note that the bounding matrix, S, is not a transition matrix since the sum of transitions out of a state will exceed unity for at least one state (except for the uninteresting case where $T = S$).

C.3. Confidence Intervals

In this section, we develop a strategy to compute a matrix S with a prescribed amount of confidence that bounds T. Let us begin by explicitly writing the transition matrix T,

$$T = \begin{bmatrix} T_{11} & \cdots & T_{1n} \\ & \vdots & \\ T_{n1} & \cdots & T_{nn} \end{bmatrix}. \tag{C.5}$$

Suppose that we observe k outputs from the process partitioned into k_i observations of state i and k_{ij} observations of transitions from state i to state j for $i, j \in \{1, \cdots, n\}$. We would like to choose a value s_{ij} to define a confidence interval $[0, s_{ij}]$: i.e., for a confidence level α our choices satisfy

$$\mathbb{P}\left[T_{ij} \leq s_{ij} | k_i, k_{ij}\right] \geq \alpha. \tag{C.6}$$

One interval with a confidence α is obtained by computing the probability that one expects to observe more transitions than were observed. Equivalently, we can define s_{ij} in terms of the observed proportion

$$s_{ij} = \min\left\{1, \frac{k_{ij}}{k_i} + \varepsilon\right\}, \tag{C.7}$$

where

$$\varepsilon = \sqrt{\frac{1}{2k_i} \log\left(\frac{1}{1-\alpha}\right)}. \tag{C.8}$$

We are now in a position to apply Hoeffding's Inequality to bound the error within the prescribed confidence.

Proposition C.2. *Let $X(t)$ be a Markov process with transition matrix T. Using the above notation, let us define the matrix S,*

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ & \vdots & \\ s_{n1} & \cdots & s_{nn} \end{bmatrix}, \tag{C.9}$$

where s_{ij} is determined by Equations C.7 and C.8. With probability $\alpha^{\min\{n^2, N\}}$, the computation of the min-entropy (as defined formally by Equation C.1) for the matrix S is a lower bound for the min-entropy of N outputs of the process:

$$H_\infty(S, \mathbf{p}, N) \leq H_\infty(T, \mathbf{p}, N). \tag{C.10}$$

A similar upper bound holds if we underestimate the initial probability distribution $\mathbf{p}$.

References

[1] Alfréd Rényi. *On measures of information and entropy.* Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960. pp. 547-561.

[2] Claude E. Shannon. *A Mathematical Theory of Communication.* Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, July, October, 1948.

[3] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, 1964, New York.

[4] Wassily Hoeffding. *Probability inequalities for sums of bounded random variables.* Journal of the American Statistical Association 58 (301), March 1963. pp. 1330.

[5] Ueli M. Maurer. *A Universal Statistical Test for Random Bit Generators.* Journal of Cryptology, Vol. 5, January 1992.