entropy

MDPI

Article

Towards More Efficient Rényi Entropy Estimation

Maciej Skorski

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 00-927 Warszawa, Poland; maciej.skorski@mimuw.edu.pl

Abstract: Estimation of Rényi entropy is of fundamental importance to many applications in cryptography, statistical inference, and machine learning. This paper aims to improve the existing estimators with regard to: (a) the sample size, (b) the estimator adaptiveness, and (c) the simplicity of the analyses. The contribution is a novel analysis of the generalized "birthday paradox" collision estimator. The analysis is simpler than in prior works, gives clear formulas, and strengthens existing bounds. The improved bounds are used to develop an adaptive estimation technique that outperforms previous methods, particularly in regimes of low or moderate entropy. Last but not least, to demonstrate that the developed techniques are of broader interest, a number of applications concerning theoretical and practical properties of "birthday estimators" are discussed.

Keywords: Rényi entropy; adaptive estimation; collision estimation; birthday paradox

Citation: Skorski, M. Towards More Efficient Rényi Entropy Estimation. *Entropy* **2023**, *25*, 185. https://doi.org/10.3390/e25020185

Academic Editors: Yong Deng and Oleg Olendski

Received: 3 November 2022
Revised: 28 December 2022
Accepted: 8 January 2023
Published: 17 January 2023

1. Introduction

1.1. Motivation and Background

The aim of *entropy estimation* is to approximately compute the Rényi entropy of an *unknown probability* distribution using only *observed samples*. Since Rényi entropy is the most established and popular uncertainty measure, the problem is not only one of fundamental interest to information theory [1], but also one of importance to a number of applied research areas. These applications of Rényi entropy include, in particular, quantifying diversity in ecology [2–4], statistical mechanics [5,6], thermodynamics [7], characterizing properties of probability distributions [8,9], DNA sequencing [10,11], network anomaly detection [12,13], clustering [14,15], data mining [16,17], predictive modelling [18], as well as security and cryptography [19–28].

To state the problem formally, if we consider a fixed discrete distribution with probability mass function p_X, the Rényi entropy [1] of some fixed positive order d is defined as

$$\mathsf{H}_d(p_X) \triangleq \frac{1}{1-d}\log_2\left(\sum_x p_X(x)^d\right). \tag{1}$$

The challenge is to estimate this quantity from n independent samples $X_1,\ldots,X_n \sim^{iid} p_X$. More precisely, we seek an *explicit function of samples* $\widehat{\mathsf{H}}$ such that the approximation

$$\widehat{\mathsf{H}}(X_1,\ldots,X_n) \approx \mathsf{H}_d(p_X) \tag{2}$$

holds with a small error, high probability, and possibly minimal sample size n. We are interested in *non-parametric estimation*, as the distribution X remains unknown.

As in prior work on Rényi entropy estimation, we focus on *integer orders*. This is not limiting, because Rényi entropies of positive integer orders: (a) encode the complete information about the distribution [29], (b) are sufficient for practical applications due to known smoothing and interpolation properties [30,31], and finally (c) are more efficient to estimate from the algorithmic perspective [32].

1.2. Related Work

The most natural, albeit not most efficient, are so-called *plugin estimators*, which insert a non-parametric estimate of the probability mass function into the entropy formula [33]. As opposed to that, sample-optimal estimators for Rényi entropy are more involved, as shown in a relatively recent line of work [32,34,35]. Loosely speaking, these estimators relate the entropy to *collision probabilities* and then take advantage of the *birthday paradox*. More specifically, base estimations are obtained by counting collisions among tuples in the observed sample, and then are optionally run in parallel to boost the statistical confidence. The birthday paradox intuitively explains why the resulting algorithms are *sublinear in the alphabet size*. Specifically, for the alphabet of size K, the sample-optimal estimation takes $O_d(K^{1-1/d})$ samples (the asymptotic notation $O_d()$ hides dependencies on d.) to achieve an additive error of at most 1 and a level of confidence of at least $2/3$ [32,34,35]. The aforesaid state-of-the-art estimators are *asymptotically minimax optimal*; that is, they achieve the asymptotically minimal sample size over the worst choice of the sampling distribution. However, their known analyses leave room for improvement with regard to simplicity, numerical precision, estimation adaptiveness, and techniques used. We elaborate on these issues below:

1. **Lack of simplicity and numerical precision.** The analyses of the state-of-the-art estimators [32,34,35] struggle with analysing the variance of collision estimators, which is tackled either by *poissonization approximations* [32,34] (which carry their own overhead) or by using *involved combinatorics* [35]. As a consequence, the variance bounds are available in asymptotic "big-Oh" notation hiding constants and higher-order dependencies (such as relations to the order d), and are not suitable for applications in statistics or cryptography, which demanding precise formulas. This point was already raised in the context of applied works on physically unclonable functions [36].
2. **Adaptiveness gap.** The focus of prior research was on establishing bounds under the *worst-case* choice over all distributions [32,34]. This is overly pessimistic because distributions that arise in practical applications are of a different structure than those occurring in this worst-case scenario analysis (the worst-case choice is known to be a mixture of uniform and Dirac distributions). The bounds were somewhat improved in the follow-up work [35] where a *prior entropy bound* is assumed. Still, there is a gap as the entropy bound is usually not known prior to the actual experiment. In fact, getting an entropy bound might be more costly than its application.
3. **Lack of established techniques.** Prior work focused on delivering asymptotic formulas and did not elaborate much on techniques that could help obtain simpler and tighter bounds. The difficulty of analyzing collision estimators is a recurring issue, well-known to the researchers working on property testing [37]. Do we have a systematic method of handling it?

1.3. Our Contribution

This work fills the aforementioned gaps with the following contributions:

1. **Simpler and more accurate analysis** of collision estimators, the main building blocks of the state-of-the-art Rényi entropy estimators. We analyze the collision estimators as *kernel averages* with the technique of *Hoeffding's decomposition*. This novelty brings the promised simplicity and improvement in accuracy.
2. **Adaptive estimation of Rényi entropy**, using no prior knowledge of the sampling distribution. The sample size cost of the presented algorithm is essentially optimal (up to a poly-logarithmic factor).
3. **Modular approach** using the established methods of *U-statistics* and *Robust Mean Estimation*. Specifically, we point out that the moment estimation problem, which Rényi entropy estimation reduces to, can be seen as the estimation of certain *U-statistics*. While the dedicated statistical theory provides the bias–variance analysis, the confidence can be independently boosted by techniques of *Robust Mean Estimation*.

This paper aims to solve the two mentioned bottlenecks and, in this way, to close the gap between the theory-oriented state-of-the-art and the demand coming from applied researchers and their practical use cases, such as [36].

1.4. Organization

The notation and preliminary concepts are discussed in Section 2. The technical results and applications are presented in Section 3. The proofs are discussed in Appendix A, and the work is concluded in Section 4.

2. Preliminaries

2.1. Basic Notation

Throughout the paper, p_X is the probability mass function of a fixed discrete distribution over an alphabet of size K and $X_1, \ldots, X_n$ are observed independent samples from p_X. We denote $[n] = \{1, \ldots, n\}$ and let $\binom{[n]}{d}$ denote the collection of all d-element subsets of n.

2.2. Estimation of Entropy, Moments, and Collisions

We leverage the following observation from prior work: the task of Rényi entropy estimation is equivalent to the task of *moment estimation*. More precisely, the d-th moment of the probability mass function p_X is defined as

$$P_d \triangleq \sum_x p_X(x)^d, \tag{3}$$

and then—immediately from definition—we have the following result:

Proposition 1. *An estimate $\widehat{H}$ of the Rényi entropy of order d (defined in (1)) has an additive error of ϵ if and only if $2^{\widehat{H}}$ is an estimate of the d-th moment (defined in (3)) with a* relative error *of $\epsilon' = 2^{\epsilon(d-1)} - 1$.*

Solving the (equivalent) problem of moment estimation is more convenient due to the beautiful representation of moments as collision probabilities. More precisely, we have

$$P_d = \mathbf{P}(X_1 = X_2 = \ldots = X_d). \tag{4}$$

2.3. U-Statistics

For a symmetric real function h of d arguments, the *U-statistic* with kernel h of the sample $X_1, \ldots, X_n$ is defined as:

$$U_h(X_1, \ldots, X_n) \triangleq \binom{n}{d}^{-1} \sum_{1 \leqslant i_1 < \ldots < i_d \leqslant n} h(X_{i_1}, \ldots, X_{i_d}). \tag{5}$$

The *U-statistic* gives an unbiased estimate of the function expectation, hence its name. *U-statistics* were invented by Hoeffding [38] to extend certain results, such as concentration bounds, to sums of partly dependent terms. Many statistical quantities can be related to *U-statistics*; for example, moments or sample variances [39]. In the same spirit, we will see estimators of the collision probability in Equation (3) as *U-statistics* and use those to establish their desired properties.

2.4. Robust Mean Estimation

It is difficult to directly obtain high confidence bounds (such as those of the Chernoff–Hoeffding type) for moment estimators. Instead, we will boost weaker bounds obtained from bias–variance analyses. To this end, we combine independent runs of estimators into high-confidence bounds using the technique of *Robust Mean Estimation*.

Estimating the mean of a distribution from i.i.d. samples is not trivial: the "obvious" use of the empirical mean is inaccurate for heavy detailed distributions. Following the recent survey [40], we mention here two solutions:

- the *median-of-means* approach organizes data (such as independent algorithm outputs) into blocks and computes the median of means within blocks.
- the *trimmed mean* approach takes the mean of independent runs, excluding a certain fraction of smallest and biggest outcomes (removing outliers).

We note that any robust mean estimation can be used to achieve confidence boosting. In this work, we stick to the median-of-means. The following result discusses its performance.

Proposition 2 (Performance of Median-of-Means [40])**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. random variables with mean μ and variance σ^2. For $k = \lceil 8 \log(1/\delta) \rceil$, split $Z_1, \ldots, Z_n$ into k blocks and let $\widehat{\mu}$ be the median of the means within blocks. Then, with probability $1 - \delta$,*

$$|\widehat{\mu} - \mu| \leqslant \sigma \sqrt{\frac{4}{\lfloor n/k \rfloor}}. \tag{6}$$

2.5. Moment Bounds

We will need some bounds on moments of probability distributions, in order to simplify formulas that arise from variance analysis. Specifically, we will use these auxiliary results to express higher-order moments in terms of moments of small order ($d = 2$ and $d = 3$).

Proposition 3. *For any probability distribution $p = (p_i)$, we have*

$$\sum_i p_i^{2d-k} \geqslant (\sum_i p_i^d)^2$$

for any integer k, d such that $1 \leqslant k \leqslant d$. Moreover, with $p_{\max} \triangleq \max_i p_i$ it holds that

$$\sum_i p_i^{2d-k} \geqslant (\sum_i p_i^d)^2 / p_{\max}^{k-1}.$$

Proposition 4. *For any non-negative sequence (p_i), it holds that the quantity $\|p\|_d \triangleq \left(\sum_i p_i^d\right)^{1/d}$ decreases in $d \geqslant 1$.*

3. Results

Following the convention from prior work, our results are stated for *moment estimation*, which is equivalent to entropy estimation as discussed in Proposition 1. Throughout the rest of the paper, we keep this reduction in mind.

3.1. Simpler & More Accurate Moment Estimation

The first novelty offered in the current work is a *simplified and strengthened* variance analysis of the state-of-the-art moment estimator, presented in Algorithm 1. Differently than in prior work, we write the estimator output as a *kernel average* of the function $h(x_{i_1}, \ldots, x_{i_d}) \triangleq \mathbf{I}(x_{i_1} = \ldots x_{i_d})$ over the d-element subsets of the sample. This approach is not only more readable but ultimately also more accurate, as it links the task of moment estimation to the established theory of *U-statistics* [38].

On top of that comes the *refined high-confidence moment estimator* in Algorithm 2, which we build on robust mean estimators [40]. Due to this modularity, it uses fewer samples than the direct approach from prior work [34].

Algorithm 1: BIRTHDAY MOMENT ESTIMATOR

Data:
- entropy order d
- samples $X_1, \ldots, X_n$, $n \geqslant d$, from an unknown distribution with p.m.f. p_X

Result: an estimate $\widehat{P}_d$ of $P_d = \sum_x p_X(x)^d$, the d-th moment of p_X

$C \leftarrow \#\left\{ \{i_1 \ldots i_d\} \in \binom{[n]}{d} : X_{i_1} = X_{i_2} = \ldots = X_{i_d} \right\}$

$\widehat{P}_d \leftarrow C / \binom{n}{d}$

return $\widehat{P}_d$

Theorem 1 (Bias–Variance Analysis of Moment Estimator)**.** *With the notation as above, the output of Algorithm 1 is unbiased:*

$$\mathbf{E}[\widehat{P}_d - P_d] = 0, \tag{7}$$

and its variance equals

$$\mathbf{Var}[\widehat{P}_d] = \frac{\sum_{k=1}^{d} \binom{d}{k}\binom{n-d}{d-k}(P_{2d-k} - P_d^2)}{\binom{n}{d}}. \tag{8}$$

In particular, for any $n \geqslant d$ and $\epsilon > 0$ we can upper-bound the variance as

$$\mathbf{Var}[\widehat{P}_d] \leqslant \frac{2d^2 P_d^{2-1/d}}{n}, \tag{9}$$

and the relative error confidence as

$$\mathbf{P}[|\widehat{P}_d - P_d| > \epsilon P_d] \leqslant \frac{2d^2}{n P_d^{1/d} \epsilon^2}. \tag{10}$$

Remark 1 (Efficient Implementation)**.** *Birthday estimators can be efficiently computed, in memory $O(n)$ and one-pass over samples, by using a dictionary to count empirical frequencies of observed elements. Such an implementation is given in the supplementary material [41].*

Remark 2 (Structural Assumptions)**.** *The bounds from Theorem 1 use the statistic P_d of the sampling distribution. This explicit dependency is beneficial, as further discussion clarifies. Lacking any prior knowledge, it can be estimated by the worst-case behavior, in terms of the alphabet size.*

Algorithm 2: HIGH-CONFIDENCE BIRTHDAY MOMENT ESTIMATOR

Data:
- a moment order d
- samples $X_1, \ldots, X_n$, $n \geqslant d$, from an unknown distribution with p.m.f. p_X
- a confidence parameter δ

Result: an estimate $\widehat{P}_d$ of $P_d = \sum_x p_X(x)^d$, the d-th moment of p_X

$k \leftarrow \lceil 8 \log(1/\delta) \rceil$

$\ell \leftarrow \lfloor n/k \rfloor$

for $j = 1, \ldots, k\rfloor$ **do**

  $\widehat{P}_d^{(j)} \leftarrow$ output of Algorithm 1 $X_{j\cdot\ell}, \ldots, X_{(j+1)\cdot\ell-1}$ on (j-th input block of length ℓ)

end

$\widehat{P}_d \leftarrow \text{MEDIAN}(\widehat{P}_d^{(j)}, j = 1, \ldots, k)$

We see that the choice $n > 6d^2\epsilon^{-2}/P_d^{1/d}$ guarantees $\mathbf{P}[|\widehat{P}_d - P_d| > \epsilon P_d] \leqslant \delta$ for $\delta = \frac{1}{3}$. Higher confidence (smaller δ) can be handled with the method of *Robust Mean Estimation*.

Theorem 2 (High-Confidence Moment Estimator)**.** *For any $\epsilon > 0$, Algorithm 2 approximates P_d with probability $1 - \delta$ and a relative error of ϵ provided that*

$$n \geqslant \left\lceil \frac{8d^2}{\epsilon^2 P_d^{1/d}} \right\rceil \cdot \lceil 8 \log(1/\delta) \rceil. \tag{11}$$

Remark 3. *The constant can be refined a little based on the methods from [42].*

When comparing these results with prior work, we review the following aspects:

1. **Novel techniques of broader interest**. We recall that analyzing variance formulas has been challenging for prior works on entropy estimation ([32,34] resorted to Poisson approximations, while [35] gave an involved combinatorial argument), even for the case $d = 2$ (the lack of sharp analysis caused lots of difficulties in property testing [37]). As opposed to these ad hoc approaches, we establish the formula in a simple yet direct manner, pointing out that such formulas can be obtained by the techniques of *U-statistics*. When discussing applications, we will further benefit from these tools.
2. **Clean and improved formulas.** Our confidence bound does not involve any implicit constants, while prior works in their main statements have unspecified dependencies on d (essentially, hiding more than absolute constants). We compare the accuracy bounds from this and prior works in Table 1 below. Our bound is strictly better given that P_d is minimized at the uniform distribution and thus $P_d^{-1/d} \leqslant K^{-1+1/d}$ (with a large gap when the distribution is far from uniform), which establishes that the dependency on d is $4d^2$. Leveraging the theory of *U-statistics*, we will show that the factor $O(d^2)$ is optimal, which is also a novel contribution.

Table 1. The performance of the "birthday estimator" of moments. In the formulas, $\epsilon \in (0, 1)$ is the relative error, n is the sample size, and $1 - \delta$ is the confidence (the prob. that $|\widehat{P}_d - P_d| \leqslant \epsilon P_d$).

Confidence $1 - \delta$ in Algorithm 1	Author	Assumption
$\delta \leqslant 4d^2 n^{-1} \epsilon^{-2} P_d^{-1/d}$	**this paper**	$n \geqslant d$
$\delta \leqslant O_d(n^{-1} \epsilon^{-2} P_d^{-1/d})$	[35]	$n \geqslant d$
$\delta \leqslant O_d(n^{-1} \epsilon^{-2} K^{1-1/d})$	[32]	$n \geqslant d$

3.2. Adaptive Estimation

As per our variance analysis, the performance actually depends not on the alphabet size K (that is ultimately the pessimistic bound) but rather on a more fine-grained statistic of X, namely P_d. Following the result in Theorem 2, we could hope for a moment estimation of

$$n =^{?} \Theta(\log(1/\delta)) \epsilon^{-2} d^2 P_d^{-\frac{1}{d}}. \tag{12}$$

The obvious obstacle is that, in general, we do not know P_d in advance. We solve this problem by developing an *adaptive algorithm*. It does not assume the right number of samples in advance but tries gradually and eventually terminates with high probability, giving the answer within the desired margin of error and using only a few more samples than the ideal bound conjectured above. Its core is a subroutine that guesses the moment value, gradually changing the candidate.

3.2.1. Lower-Bounding Moments

The key ingredients of our approach are the following two subroutines: Algorithm 3 tests, based on samples, whether the moment is smaller or bigger than a proposed candidate; subsequently, Algorithm 4 loops the tester over a grid of candidate values.

The correctness of the approach is guaranteed by the lemmas stated below.

Algorithm 3: MOMENTLESSTHAN

Data:
- independent samples $X_1, \ldots, X_n \sim^{iid} p_X$
- tested threshold Q
- access to an estimator $\widehat{P}$ of P_d

Result: tests if $P_d \leqslant \frac{Q}{2}$ or $P_d \geqslant 2Q$
$\widehat{P} \leftarrow \widehat{P}(X_1, \ldots, X_n)$
if $\widehat{P} < Q$ **then**
  return `True`
else if $\widehat{P} \geqslant Q$ **then**
  return `False`

Algorithm 4: MOMENTBOUND

Data:
- independent samples $X_1, X_2, \ldots \sim^{iid} p_X$ (online access)
- access to Algorithm 3
- the sample size $n(Q, \delta)$ for Algorithm 3

Result: finds bound Q such that $\frac{Q}{2} \leqslant P_d \leqslant Q$
$Q \leftarrow 1$
while $Q \geqslant 1/K^{d-1}$ & $b =$ `True` **do**
  $Q \leftarrow Q/2$
  $n \leftarrow n(Q, \delta/\log K)$
  $b \leftarrow$ MOMENTLESSTHAN$(X_1, \ldots, X_n; Q)$ `/* this call can recycle` X_i `*/`
end
return Q

Lemma 1 (Moment Testing). *Let $\widehat{P}$ be any estimator of P_d which, given*

$$n \geqslant C(\delta) P_d^{-1/d} \epsilon^{-2}$$

samples, for some function $C(\cdot)$ and any $\epsilon > 0, 1 > \delta > 0$, achieves a relative error of ϵ with probability $1 - \delta$. Then Algorithm 3, when given at least

$$n(Q, \delta) = \lceil 4C(\delta) Q^{-1/d} \rceil$$

samples of X, with probability $1 - \delta$, outputs TRUE *when $P_d \leqslant Q/2$ and* FALSE *when $P_d \geqslant 2Q$.*

Lemma 2 (Moment Bounding). *With same $\widehat{P}$ as in Lemma 1, with probability $1 - \delta$, Algorithm 4 terminates after using at most*

$$n = \lceil 4C(\delta/(d-1)\log K) P_d^{-1/d} \rceil$$

samples of X, and its output Q satisfies $\frac{Q}{2} \leqslant P_d \leqslant 4Q$.

3.2.2. Construction of Adaptive Estimator

Armed with Lemma 2, we are ready to analyze adaptive estimation. The algorithm is the same in the non-adaptive case, and we need Lemma 2 only to adjust the sample size.

Theorem 3 (Adaptive Moment Estimation). *With online access to samples from X, with probability $1 - \delta$, one can estimate P_d within a relative error of ϵ, terminating at the number of samples*

$$n \leqslant O((\log(1/\delta) + \log\log K) P_d^{-\frac{1}{d}} \epsilon^{-2}). \tag{13}$$

Remark 4 (Adaptive Overhead). *The sample size for adaptive estimation differs from the "dream bound" in (12) by a (very small) factor of* $\log\log K$.

Remark 5 (Sample Complexity Guarantees). *Observe that the algorithm is not guaranteed to achieve the "good" sample complexity every time, but rather with high probability. This is a minor issue inherently related to the concept of adaptive estimation and does not affect much the performance in practical applications. Namely, we can always clip the total number of samples available at the pessimistic level from prior work and fall back to the fixed-size sample estimation should the adaptive estimation exceed the limit. This, however, happens with small probability δ, which can be further decreased with little overhead.*

Remark 6 (Comparison to Prior Work). *We give a clear comparison of our adaptive estimation and prior work in Table 2 below. We always have $P_d \geqslant Q$ and $P_d \geqslant K^{-1+1/d}$. Furthermore, usually P_d/Q is much bigger than 1 (because in practice we do not know a prior lower bound in advance) and P_d is much bigger than $K^{-1+1/d}$ (this gap can be as big as $K^{\Omega(1)}$, which is a huge factor for some applications; for example, in cryptography, we consider alphabets as big as $K = 2^{256}$). Thus, our bound outperforms the previous approaches for typical use cases.*

Table 2. The performance of moment (Rényi entropy) estimators. In the formulas, K is the alphabet size, ϵ is the relative error, and the confidence is $1 - \delta$.

Sample Size n	Author	Assumptions
$O((\log(1/\delta) + \log\log K)\epsilon^{-2}P_d^{-1/d})$	**this paper**	
$O(\log(1/\delta)\epsilon^{-2}Q^{-\frac{1}{d}})$	[35]	prior bound $Q \leqslant P_d$
$O(\log(1/\delta)\epsilon^{-2}K^{1-\frac{1}{d}})$	[32]	

3.3. Novel Applications

3.3.1. One-Sided Estimation: Random Sources for Cryptography

The collision probability P_d for $d = 2$ plays an important role in cryptography: it quantifies the amount of randomness that can be extracted from a distribution [28]. For that extraction to work, P_2 should be small enough. Specifically, if $P_2 \leqslant 2^{-k}$ allows for extraction of nearly k bits of cryptographic quality, how could we check whether $P_2 \leqslant 2^{-k}$?

To solve the problem, we adapt Algorithm 4 by adding early stopping; namely, we quit the loop if $Q \geqslant 2^{-k}$. We take $n = O(\log(k/\delta)2^{k/2})$ samples.

It remains for us to show that the algorithm behaves as desired. By the guarantees in Lemma 1 and the union bound over at most k steps, with probability $1 - \delta$, we have the following: when $P_2 < 2^{-k-1}$, the algorithm finishes with $Q \leqslant 2^{-k}$; and when $P_2 > 2^{-k+1}$, the algorithm finishes with $Q \geqslant 2^{-k}$. This can be generalized to one-sided estimation for any d, where the goal is to decide whether $P_d > \Omega(2^{-k})$ or $P_d < O(2^{-k})$.

This one-sided estimation allows for saving samples and testing only up to a necessary extent. In cryptography, we do not have to estimate the whole entropy (which may be more costly, even with adaptive estimation) but only what suffices for the chosen application.

3.3.2. Birthday Estimators Are UMVUE

The shortcut UMVUE stands for *uniformly minimum unbiased variance estimators*. We prove this conceptually strong and interesting characterization, which essentially shows that the birthday estimator in Theorem 1 is variance-optimal among unbiased estimators. The argument is inspired by our variance analysis, seeing the estimator as a *U-statistic*.

Corollary 1 (Birthday estimators are UMVUE). *Let $\widehat{P}_d$ be as in Algorithm 1 and $\tilde{P}_d$ be another unbiased (for any X) estimator of P_d. Then we have $\mathbf{Var}[\widehat{P}_d] \leqslant \mathbf{Var}[\tilde{P}_d]$.*

Proof. We will use the known result due to Lehmann and Sheff, which states that if an unbiased estimator is a function of a complete and sufficient data statistic, then it has the smallest possible variance [43,44].

To apply this result, without losing generality (as it is a matter of encoding the alphabet), we assume that X takes values in the set $\{1, \ldots, K\}$. Consider the sample $X_1, \ldots, X_n$, and let σ be the rearrangement such that $X_{\sigma(1)} \leqslant X_{\sigma(2)} \leqslant \ldots X_{\sigma(n)}$ (this is called the *order statistic*). The estimator $\widehat{P}_d$ can be seen as the average of the symmetric function $h(X_{i_1}, \ldots, X_{i_d}) = \mathbf{I}(X_{i_1} = x_{i_2} = \ldots = X_{i_d})$ over tuples $i_1 < i_2 < \ldots < i_d$, and thus is also the function of $T = (X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)})$. The claim follows if we prove that T is sufficient and complete (as a sample statistic).

Order statistics are sufficient for univariate distributions. This is because we have:

$$P_{X_1,\ldots,X_n | X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}} = \frac{1}{n!},$$

which does not depend on X_i. Thus, $X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}$ carries the same information about the data as X_i.

The completeness of T means that there are no non-trivial unbiased estimators of zero; equivalently, if $\mathbf{E}f(T) = 0$ for all sampling distributions and some function f, then $\mathbf{P}[f(T) = 0] = 1$. To this end, observe that from the sufficiency proved above we have

$$\mathbf{E}f(X_{\sigma(1)}, \ldots, X_{\sigma(n)}) = 0 \Longrightarrow \mathbf{E}f(X_1, \ldots, X_n) = 0$$

Suppose that the above holds for any finitely supported sampling distribution X. Let X take values $i \in I$ with probability p_i. Then the above implies

$$\sum_{i_1,\ldots,i_n \in I \times \ldots \times I} p_{i_1} \ldots p_{i_n} f(i_1, \ldots, i_n) = 0$$

for every distribution (p_i). The left-hand side represents a multivariate polynomial in variable p_i, which evaluates to zero on the entire simplex of dimension $n - 1$. Thus, its coefficients must be zero, which implies $f(i_1, \ldots, i_n) = 0$ for each tuple $i_1, \ldots, i_n$ and proves that T is sufficient. □

3.3.3. Central Limit Theorem for Birthday Estimators

We again represent the estimator as the average sum over tuples:

$$\widehat{P}_d = \binom{n}{d}^{-1} \sum_{i_1 < \ldots < i_d} h(X_{i_1}, \ldots, X_{i_d}),$$

where

$$h(x_{i_1}, \ldots, x_{i_d}) \triangleq \mathbf{I}(x_{i_1} = x_{i_2} = \ldots = x_{i_d}).$$

We view the whole expression as the *U-statistic* with the kernel function h. Then we show the following strong result (below, $\mathcal{N}(0, \sigma^2)$ denotes the normal distribution with zero-mean and variance σ^2).

Corollary 2 (Asymptotic Normality)**.** *For $n \to +\infty$ it holds that $\sqrt{n} \cdot (\widehat{P}_d - P_d) \to \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = d^2(P_{2d-1} - P_d^2)$, with P_d as in (3).*

The proof utilizes the classical convergence results for *U-statistics* [38] and the derivation of our variance formula. Note that the result says that the central limit theorem applies, despite the fact that the sum components are correlated. Clearly, the result is interesting on its own, particularly because (a) it proves that our constant $O(d^2)$ is sharp, and (b) can be used more generally to benchmark other proposed bounds, by means of comparing with the asymptotic gaussian tail.

However, we would like to point out an application to applied statistical research. In [36], Rényi entropy of order $d = 2$ has been estimated for the distribution of physically

unclonable functions (PUFs), which are important in the field of cryptography. However, their methodology lacks statistical rigor. Particularly, for the authors' needs, prior work on Rényi entropy estimation was insufficient in terms of clarity on constants; thus, they resorted to the naive application of the central limit theorem, which can give very biased results.

A more solid alternative would be to use the above corollary to (a) justify the soundness, at least in the regime of large n, and (b) establish a more robust estimation of the variance.

Proof of Corollary 2. The limiting variance equals $d^2\sigma_1^2$ [39] with

$$\sigma_1^2 = \mathbf{Cov}[h(X_{i_1}, \ldots, X_{i_d}), h(X_{j_1}, \ldots, X_{j_d})],$$

where the tuples $(i_1, \ldots, i_d)$ and $(j_1, \ldots, j_d)$ have only $k = 1$ element in common. We analyze this expression when proving Theorem 1 and know that it equals $P_{2d-1} - P_d^2$. The claimed formula now follows. □

3.3.4. Adaptive Testing in Evaluation of PUFs

Here we discuss again an application to [36], but from a different perspective. As explained by the authors, the problem with estimating Rényi entropy of PUFs is a serious bottleneck: for this problem, the alphabet is huge, which limits the experiment scope, even on computational clusters [36]. In this note, we would like to point out that parts of these difficulties can be solved by our adaptive estimation. In fact, PUFs provide an excellent use case when *entropy is quite low*; therefore, the moment P_d term in Theorem 3 is much bigger than the pessimistic bound based on the alphabet size. We discuss this application in full detail in a follow-up work.

3.3.5. Applications to Property Testing

The estimator from Algorithm 1 was first studied in [45], but the variance bounds obtained were not sharp. Quite oddly, in the ongoing research on closeness testing, the birthday-like collision estimators (being subroutines for uniformity checking) seemed to be suboptimal [46] until, very recently, the work of [37] re-examined the variance formula for $d = 2$ and shows that it achieves (in our notation) optimal dependence on K and ϵ. Thus, a breakthrough was possible just because of a specialized version of (8). In this discussion, we would like to (a) point out that the general variance formula can likely have similar applications and impact for $d > 2$ and should be of broader interest, and (b) comment on a minor gap in an early version of the proof of the central result in [37]. Lemma 2.3 in [37], which establishes the variance bound for $d = 2$, is the key ingredient of the main results. The authors derive an expression bounding, in our notation, the variance in Theorem 1 for the case $d = 2$. When doing so, they face up the term $n(n-1)(n-2)(P_3 - P_2^2)$ (in our notation) and bound it as $n^3(P_3 - P_2^2)$ (the last line of derivation claims the upper bound, the following remark claims the lower bound). The reasoning, however, is valid when $P_3 - P_2^2$ is non-negative. This is true by Proposition 3.

3.4. Application to Statistical Inference

We will use Algorithm 1 to efficiently test whether a given distribution is close or far from a uniform one. The procedure described below is asymptotically equivalent but numerically superior to the one from [37].

Denote by K the alphabet size and let γ be such that $P_2 = \frac{1}{K} + \gamma$. We see that

$$\gamma = \sum_x \left(p_X(x) - \frac{1}{K} \right)^2, \tag{14}$$

which shows that γ measures the squared ℓ_2-distance between p_X and the uniform distribution. For convenience, we will refer to γ as the *collision gap*. Define

$$\widehat{\gamma} = \widehat{P}_2 - \frac{1}{K}. \tag{15}$$

This estimator gives an unbiased approximation of γ, because $\mathbf{E}[\widehat{\gamma}] = \mathbf{E}[\widehat{P}_2] - \frac{1}{K} = P_2 - \frac{1}{K} = \gamma$. Furthermore, its variance equals the variance of $\widehat{P}_2$ because $\frac{1}{K}$ is a deterministic constant. By Chebyshev's inequality, with probability $1 - \delta$, we have

$$|\widehat{\gamma} - \gamma| \leqslant \sqrt{\mathbf{Var}[\widehat{P}_2]/\delta}. \tag{16}$$

We now analyze the variance in more detail. Define $\delta(x) = p_X(x) - \frac{1}{K}$. Then $P_2 = \frac{1}{K} + \sum_x \delta(x)^2$ and $P_3 - P_2^2 = \sum_x \delta(x)^3 + \frac{1}{K}\sum_x \delta(x)^2 - (\sum_x \delta(x)^2)^2$. In particular, we have $P_2 = \frac{1}{K} + \gamma$ and $0 \leqslant P_3 - P_2^2 \leqslant \gamma^{\frac{3}{2}} + \frac{\gamma}{K}$ by Propositions 3 and 4. Therefore, by Theorem 1 we obtain

$$\mathbf{Var}\left[\widehat{P}_d - \frac{1}{K}\right] \leqslant \frac{4(\frac{\gamma}{K} + \gamma^{\frac{3}{2}})}{n} + \frac{\gamma + \frac{1}{K}}{\binom{n}{2}}. \tag{17}$$

Thus, with probability $1 - \delta$, we have

$$\gamma - \sqrt{\frac{4(\frac{\gamma}{K} + \gamma^{\frac{3}{2}})}{n\delta} + \frac{\gamma + \frac{1}{K}}{\binom{n}{2}\delta}} \leqslant \widehat{\gamma} \leqslant \gamma + \sqrt{\frac{4(\frac{\gamma}{K} + \gamma^{\frac{3}{2}})}{n\delta} + \frac{\gamma + \frac{1}{K}}{\binom{n}{2}\delta}}. \tag{18}$$

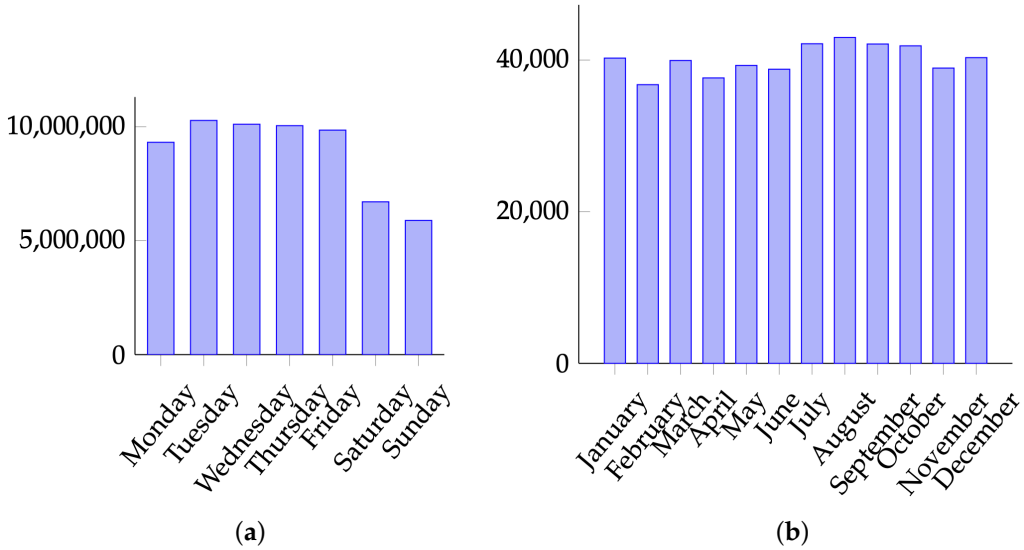
The above two-sided inequality allows us to estimate a range of possible values γ with respect to the (observed) statistics $\widehat{\gamma}$, which yields high-confidence bounds for γ.

We illustrate the procedure numerically on real-world datasets. Data and results are summarized in Figure 1a,b. Our method confirms non-uniformity in both cases and provides confidence intervals. The details of the experiment are shared in the supplementary material [41].

3.5. Application to Entropy Estimation

The following experiment illustrates advantages of adaptive entropy estimation for distributions with large support and relatively low entropy, such as Zipf's law.

Let X follow Zipf's law with parameter $s = 1.1$ and the support of $K = 10^4$ elements. By numerical calculations, we find that $P_2 \approx 0.40005$. Consider now the task of estimating entropy of X from samples. Theorem 3 allows us to save a large factor of about $K^{1/2} = 10^2$ in the number of samples. Calculations show that on a sample of size about $n = 10,000$, the algorithm from Algorithm 2 finds an approximation $\widehat{P}_d = 0.39898$ with a relative error $\epsilon = \frac{1}{2}$ and confidence $1 - \delta = 0.95$. The details appear in the supplementary material [41].



(**a**) (**b**)

Figure 1. (**a**) U.S. births 2000–2014 (source: Social Security Administration). For this dataset, $\widehat{P}_2 \approx 0.14794$ and $\frac{1}{K} \approx 0.14286$, so that the gap equals $\gamma \approx 0.00017$. Our method gives the 99% confidence interval of (0.005,0.00516). (**b**) Births from insurance claims (source: courtesy of Roy Murphy). For this dataset, $\widehat{P}_2 \approx 0.08348$ and $\frac{1}{K} \approx 0.08333$, so that the collision gap equals $\gamma \approx 0.00015$. Our method gives the 99% confidence interval of (0.00009,0.00035).

4. Conclusions

This work simplifies the variance analysis of collision estimators, establishing the closed-form exact formulas and improving upon prior data-oblivious bounds by making them dependent on certain data statistics. In particular, we use the derived formulas to estimate Rényi entropy adaptively, asymptotically, and give other applications.

Numerical experiments highlight the importance of the dependency of sample size on confidence. The constants involved exponentially affect the confidence, so that further improvements are of significance for many real-world inference problems. This problem is left for future research.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Data Availability Statement: The data and code is shared in the GitHub repository [41].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Appendix A.1. Proof of Proposition 3

For convenience, we introduce the function $f(u) = u^{d-1}$. Then we have:

$$\sum_i p_i f(p_i)^2 \geqslant (\sum_i p_i f(p_i))^2,$$

which is easiest to see by writing equivalently $\mathbf{Var}[Z] \geqslant 0$ where Z is a random variable taking values p_i with probabilities p_i. We therefore have

$$\sum_i p_i^{2d-1} \geqslant (\sum_i p_i^d)^2.$$

Since $0 \leqslant p_i \leqslant 1$ and $1 \leqslant k \leqslant d$, we have $p_i^{2d-1} \leqslant p_i^{2d-k}$. We finally obtain

$$\sum_i p_i^{2d-k} \geqslant \sum_i p_i^{2d-1} \geqslant (\sum_i p_i^d)^2.$$

This can be further sharpened using $p_i^{2d-1} \leqslant p_i^{2d-k} p_{\max}^{k-1}$, which gives

$$\sum_i p_i^{2d-k} p_{\max}^{k-1} \geqslant \sum_i p_i^{2d-1} \geqslant (\sum_i p_i^d)^2.$$

Appendix A.2. Proof of Proposition 4

The property in Proposition 4 is equivalent to the monotonicity of p-norms. For a proof of the latter one, see, for example [47].

Appendix A.3. Proof of Theorem 1

Appendix A.3.1. Proof of the Bias–Variance Formula

For a tuple index $\mathbf{i} = (i_1, \ldots, i_d)$ where components $i \in \mathbf{i}$ are different, let $C_\mathbf{i}$ indicate whether all X_i for $i \in \mathbf{i}$ collides. The definition of the estimator directly implies that

Proposition A1 (Mean Value Formula)**.** *The output of Algorithm 1 is*

$$\widehat{P}_d = \binom{n}{d}^{-1} \sum_{\mathbf{i}=(i_1,\ldots,i_d):1\leqslant i_1<i_2<\ldots<i_d\leqslant n} C_\mathbf{i}. \tag{A1}$$

From Proposition A1, it is straightforward to see that the estimator is unbiased.

Proposition A2 (Estimator Is Unbiased)**.** *For any* $\mathbf{i}$ *we have that* $\mathbf{E}C_\mathbf{i} = \sum_x p_X(x)^d = P_d$*. In particular, the estimator* $\widehat{P}_d$ *of the d-moment is unbiased.*

In order to establish a variance formula, we analyze the covariance of the terms $C_\mathbf{i}$.

Proposition A3 (Covariance Formulas)**.** *Let* $\mathbf{i} = i_1, \ldots, i_d$ *and* $\mathbf{j} = j_1, \ldots, j_d$ *be tuples of distinct indices. Suppose that exactly* $k \geqslant 0$ *of entries in* $\mathbf{i}$ *collides with some entries in* $\mathbf{j}$*, that is* $|\mathbf{i} \cap \mathbf{j}| = k$*. Then we have*

$$\mathbf{E}\left[C_\mathbf{i} C_\mathbf{j}\right] = \begin{cases} (\sum_x p_X(x)^d)^2 & k = 0 \\ \sum_x p_X(x)^{2d-k} & k > 0, \end{cases}$$

which implies that

$$\mathbf{Cov}[C_\mathbf{i}, C_\mathbf{j}] = \begin{cases} P_{2d-k} - P_d^2 & k = |\mathbf{i} \cap \mathbf{j}| > 0 \\ 0 & |\mathbf{i} \cap \mathbf{j}| = 0. \end{cases}$$

Remark A1 (Overlaps imply positive correlation)**.** *Note that* $\mathbf{E}\left[C_\mathbf{i} C_\mathbf{j}\right] - \mathbf{E}\left[C_\mathbf{i}\right]\mathbf{E}\left[C_\mathbf{j}\right]$ *equals* $\sum_x p_X(x)^{2d-k} - (\sum_x p_X(x)^d)^2$*. For* $k \leqslant d-1$*, we have* $\sum_x p_X(x)^{2d-k} \geqslant \sum_x p_X(x)^{d+1}$ *and* $\sum_x p_X(x)^{2d-1} = \sum_x p_X(x) p_X(x)^{2(d-1)} \geqslant (\sum_x p_X(x)^d)^2$ *by Jensen's inequality. Thus, we have positive correlation. This can also be shown by the FKG correlation inequality [48].*

Proof. We first prove the formula for $\mathbf{E}\left[C_\mathbf{i} C_\mathbf{j}\right]$. Consider the case $k = 0$, which means that $\mathbf{i}$ and $\mathbf{j}$ do not share a common index; it is easy to see that the formula is true because $2d$ random variables X_i, X_j for $i \in \mathbf{i}$ and $j \in \mathbf{j}$ are independent (lack of collisions among the indices) so that $\mathbf{E}[C_\mathbf{i} C_\mathbf{j}] = \mathbf{E}[C_\mathbf{i}]\mathbf{E}[C_\mathbf{j}]$. Consider now the case $k > 0$, which means that $\mathbf{i}$ and

$\mathbf{j}$ overlap. We have $X_i = X_j$ for all $i, j \in \mathbf{i} \cap \mathbf{j}$. Conditioning on this common value of all X_i and X_j (call it x) and denoting $X_{\mathbf{i}} = (X_i)_{i \in \mathbf{i}}$, we obtain

$$\mathbf{E}\left[C_{\mathbf{i}} C_{\mathbf{j}} \middle| X_{\mathbf{i}} = X_{\mathbf{j}} = \underbrace{x, \ldots, x}_{2d-k} \right] = p_X(x)^{2d-k},$$

because we have exactly $2d - k$ distinct variables X_i or X_j and all are equal to x. The formula follows now by aggregating over possible values of x.

The covariance formula follows now by combining the above bounds and Proposition A1, because $\mathbf{Cov}[C_{\mathbf{i}}, C_{\mathbf{j}}] = \mathbf{E}[C_{\mathbf{i}} C_{\mathbf{j}}] - \mathbf{E}[C_{\mathbf{i}}]\mathbf{E}[C_{\mathbf{j}}]$. □

Finally, the covariance bounds from Proposition A1 and A3 remain to be used. We need the following fact, which counts how many times we see a particular pattern from the covariance formula.

Proposition A4 (Number of terms)**.** *There are* $\binom{n}{d}\binom{d}{k}\binom{n-d}{d-k}$ unordered *distinct tuples* $\mathbf{i}$ *and* $\mathbf{j}$ *that satisfy* $|\mathbf{i} \cap \mathbf{j}| = k$*. The number of ordered tuples equals* $\binom{n}{2d-k}$*.*

Proof. Recall that $\mathbf{i}$ and $\mathbf{j}$ are d-combinations out of n. To enumerate tuples such that $|\mathbf{i} \cap \mathbf{j}| = k$, note that it suffices to choose $\mathbf{i}$ one in $\binom{n}{d}$ ways, then choose k common elements in $\binom{d}{k}$ ways and then choose remaining $\mathbf{j} \setminus \mathbf{i}$ elements in $\binom{n-d}{d-k}$ ways. This gives the formula. □

By Proposition A1 we have

$$\mathbf{Var}[\widehat{P}_d] = \binom{n}{d}^{-2} \sum_{\mathbf{i},\mathbf{j}:\text{distinct ordered } d\text{-tuples}} \mathbf{Cov}[C_{\mathbf{i}}, C_{\mathbf{j}}].$$

Now Equation 8 follows by using Propositions A3 and A4.

Appendix A.3.2. Proof of Variance Upper Bound (9)

Observe that

$$\mathbf{Var}[\widehat{P}_d] \leqslant \frac{\sum_{k=1}^{d} \binom{d}{k}\binom{n-d}{d-k} P_{2d-k}}{\binom{n}{d}}.$$

Since we have $P_{2d-k} \leqslant P_d^{\frac{2d-k}{d}}$ by Proposition 4, we can estimate

$$\mathbf{Var}[\widehat{P}_d] \leqslant \frac{P_d^2 \sum_{k=1}^{d} \binom{d}{k}\binom{n-d}{d-k} P_d^{-k/d}}{\binom{n}{d}}.$$

Let us rewrite the formula as follows

$$\sum_{k=1}^{d} \binom{d}{k}\binom{n-d}{d-k} P_d^{-k/d} = \sum_{k=1}^{d} A_k,$$

where we define

$$A_k \triangleq \binom{d}{k}\binom{n-d}{d-k} P_d^{-k/d}.$$

By the definition of binomial coefficients

$$\frac{A_{k+1}}{A_k} = \frac{d-k}{k+1} \cdot \frac{d-k}{n-2d+k+1} \cdot \frac{1}{P_d^{-k/d}}.$$

Since the ratio, with respect to $k = 1 \dots d$, is maximized when $k = 1$ (because it decreases in k), we obtain:

$$\frac{A_{k+1}}{A_k} \leqslant \frac{1}{2} \quad \text{when } \frac{d-1}{2} \cdot \frac{d-1}{n-2d+2} \cdot \frac{1}{P_d^{-1/d}} \leqslant \frac{1}{2}.$$

The condition above is equivalent to $n - 2d + 2 \geqslant (d-1)^2 / P_d^{1/d}$, and holds when

$$n \geqslant 2d^2 / P_d^{1/d}. \tag{A2}$$

Indeed, since $0 < P_d \leqslant 1$, we obtain $n \geqslant d^2 / P_d^{1/d} \geqslant ((d-1)^2 + 2d - 2) P_d^{1/d} \geqslant (d-1)^2 P_d^{1/d} + 2d - 2$. Thus:

$$\sum_{k=1}^{d} A_k \leqslant 2A_1,$$

and consequently:

$$\mathbf{Var}[\widehat{P}_d] \leqslant \frac{2d\binom{n-d}{d-1} P_d^{2-1/d}}{\binom{n}{d}}.$$

Finally, we can observe that

$$\begin{aligned} \frac{d\binom{n-d}{d-1}}{\binom{n}{d}} &= d^2 \cdot \frac{(n-d)!}{(n-2d+1)!} \cdot \frac{(n-d)!}{n!} \\ &= d^2 \cdot \frac{(n-d)^{\underline{d-1}}}{n^{\underline{d}}} \\ &\leqslant \frac{d^2}{n}, \end{aligned}$$

so that we get the variance bound:

$$\mathbf{Var}[\widehat{P}_d] \leqslant 2d^2 / n \cdot P_d^{2-1/d}, \tag{A3}$$

and the confidence bound:

$$\mathbf{P}[|\widehat{P}_d - P_d| > \epsilon P_d] \leqslant \frac{2d^2}{n\epsilon^2 P_d^{1/d}} \tag{A4}$$

This bound for $\epsilon < 1$ is also meaningful when $n < 2d^2 / P_d^{1/d}$ because the probability is at most 1. This completes the proof.

Appendix A.3.3. Proof of Theorem 2

The result follows by combining Theorem 1 and Proposition 2.

Appendix A.4. Proofs of Adaptive Testing

Appendix A.4.1. Proof of Lemma 1

Suppose now that $P_d \leqslant Q/2$. Then we have that

$$Q \geqslant P_d + Q/2, \tag{A5}$$

and thus

$$\mathbf{P}[\widehat{P} \leqslant Q] \geqslant \mathbf{P}[\widehat{P} - P_d \leqslant Q/2]$$

Let $\epsilon = Q/2P_d$, then $C(\delta)P_d^{-1/d}\epsilon^{-2} = 4C(\delta)P_d^{2-1/d}Q^{-2} \leqslant C(\delta)(Q/2)^{-1/d} \leqslant n$. Therefore, $\mathbf{P}[\widehat{P} - P_d \leqslant Q/2] = \mathbf{P}[\widehat{P} \leqslant P_d + \epsilon P_d] \geqslant 1 - \delta$ so that we conclude

$$\mathbf{P}[\widehat{P} \leqslant Q] \geqslant \mathbf{P}[\widehat{P} - P_d \leqslant \epsilon P_d] \geqslant 1 - \delta.$$

Suppose that $P_d \geqslant 2Q$. Then we have that

$$Q \leqslant P_d - P_d/2,$$

which implies

$$\mathbf{P}[\widehat{P} \geqslant Q] \geqslant \mathbf{P}[\widehat{P} - P_d \geqslant -P_d/2].$$

Setting $\epsilon = \frac{1}{2}$, we obtain $C(\delta)P_d^{-1/d}\epsilon^{-2} \leqslant 4C(\delta)(2Q)^{-1/d} \leqslant n$, and so

$$\mathbf{P}[\widehat{P} \leqslant Q] \geqslant \mathbf{P}[\widehat{P} - P_d \geqslant -\epsilon P_d] \geqslant 1 - \delta,$$

which finishes the proof.

Appendix A.4.2. Proof of Lemma 2

Observe that the internal loop can do at most $(d-1)\log K$ steps. This is because we decrease the candidate bound Q in each step by a factor of 2, starting from $Q = 1$, down to the smallest possible value of $\frac{1}{K^{d-1}}$ (the smallest possible moment value, achieved by the uniform distribution). Therefore, when we set $\delta = \delta/(d-1)\log K$ in the subroutine, the guarantees from Lemma 1 hold in every step. Suppose that the loop takes exactly k steps. This means that the output is $Q = 2^{-k}$ and that the subroutine outputs $b = \textsc{True}$ at step $k-1$, which gives $P_d \leqslant 2^{-k+2} = 4Q$ by Lemma 1; at step k, we either get $\textsc{False}$, which means $P_d \geqslant 2^{-k-1} = Q/2$ or $2^{-k} \leqslant K^{-d+1}$ and then $P_d \geqslant K^{-d+1} \geqslant 2^{-k} > 2^{-k-1} = Q/2$.

We shall also clarify how the online sample access is used: as stated in the algorithm, we recycle X_i samples already so that we request only "missing" samples when the number of samples n increases due to the change of the candidate Q. This recycling is captured in the union bound.

Appendix A.5. Proof of Theorem 3

We use $\widehat{P}$ from Theorem 1 combined with the *Robust Mean Estimation*, which works with accuracy ϵ and confidence $1 - \delta$ given the number of samples, as in Theorem 2. This shows that $n = \Theta(\log(1/\delta)\epsilon^{-2}P_d^{-1/d})$ works. Now it suffices to combine this with Lemma 2.

References

1. Rényi, A. On measures of information and entropy. In Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability , Berkeley, CA, USA, 20 June–30 July 1960; pp. 547–561.
2. Masisi, L.; Nelwamondo, V.; Marwala, T. The use of entropy to measure structural diversity. In Proceedings of the 2008 IEEE International Conference on Computational Cybernetics, Stara Lesna, Slovakia, 27–29 November 2008; pp. 41–45. [CrossRef]
3. Lövei, G. Generalised entropy indices have a long history in ecology—A comment. *Community Ecol.* **2005**, *6*, 245–247. [CrossRef]
4. Leinster, T.; Meckes, M. Maximizing Diversity in Biology and Beyond. *Entropy* **2016**, *18*, 88. [CrossRef]
5. Lenzi, E.; Mendes, R.; da Silva, L. Statistical mechanics based on Renyi entropy. *Phys. A Stat. Mech. Its Appl.* **2000**, *280*, 337–345. [CrossRef]
6. Ansari, M.H.; Nazarov, Y.V. Exact correspondence between Renyi entropy flows and physical flows. *Phys. Rev. B* **2015**, *91*, 174307. [CrossRef]
7. Czinner, V.G.; Iguchi, H. Rényi entropy and the thermodynamic stability of black holes. *Phys. Lett. B* **2016**, *752*, 306–310. [CrossRef]
8. Fashandi, M.; Ahmadi, J. Characterizations of symmetric distributions based on Rényi entropy. *Stat. Probab. Lett.* **2012**, *82*, 798–804. [CrossRef]
9. Golshani, L.; Pasha, E. Rényi entropy rate for Gaussian processes. *Inf. Sci.* **2010**, *180*, 1486–1491. [CrossRef]
10. Vinga, S.; Almeida, J.S. Rényi continuous entropy of DNA sequences. *J. Theor. Biol.* **2004**, *231*, 377–388. [CrossRef]

11. Vinga, S.; Almeida, J.S. Local Renyi entropic profiles of DNA sequences. *BMC Bioinform.* **2007**, *8*, 393. [CrossRef]
12. Li, K.; Zhou, W.; Yu, S.; Dai, B. Effective DDoS Attacks Detection Using Generalized Entropy Metric. In Proceedings of the Algorithms and Architectures for Parallel Processing, 9th International Conference, ICA3PP 2009, Taipei, Taiwan, 8–11 June 2009; pp. 266–280. [CrossRef]
13. Bereziński, P.; Jasiul, B.; Szpyrka, M. An Entropy-Based Network Anomaly Detection Method. *Entropy* **2015**, *17*, 2367–2408. [CrossRef]
14. Liang, J.; Zhao, X.; Li, D.; Cao, F.; Dang, C. Determining the number of clusters using information entropy for mixed data. *Pattern Recognit.* **2012**, *45*, 2251–2265. [CrossRef]
15. Jenssen, R.; Hild, K.; Erdogmus, D.; Principe, J.; Eltoft, T. Clustering using Renyi's entropy. In Proceedings of the International Joint Conference on Neural Networks, Portland, OR, USA, 20–24 July 2003; Volume 1, pp. 523–528. [CrossRef]
16. Cornforth, D.J.; Jelinek, H.F. Detection of congestive heart failure using Renyi entropy. In Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September 2016; pp. 669–672.
17. Cornforth, D.J.; Tarvainen, M.P.; Jelinek, H.F. Using renyi entropy to detect early cardiac autonomic neuropathy. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 5562–5565. [CrossRef]
18. Gajowniczek, K.; Ząbkowski, T.; Orłowski, A. Comparison of Decision Trees with Rényi and Tsallis Entropy Applied for Imbalanced Churn Dataset. *Ann. Comput. Sci. Inf. Syst.* **2015**, *5*, 39–44. [CrossRef]
19. Knuth, D.E. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*; Addison Wesley Longman Publishing Co., Inc.: Redwood City, CA, USA, 1998.
20. van Oorschot, P.C.; Wiener, M.J. Parallel Collision Search with Cryptanalytic Applications. *J. Cryptol.* **1999**, *12*, 1–28. [CrossRef]
21. Arikan, E. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Inf. Theory* **1996**, *42*, 99–105. [CrossRef]
22. Pfister, C.E.; Sullivan, W. Rényi Entropy, Guesswork Moments, and Large Deviations. *IEEE Trans. Inf. Theory* **2004**, *50*, 2794–2800. [CrossRef]
23. Hanawal, M.K.; Sundaresan, R. Guessing Revisited: A Large Deviations Approach. *IEEE Trans. Inf. Theory* **2011**, *57*, 70–78. [CrossRef]
24. Impagliazzo, R.; Zuckerman, D. How to Recycle Random Bits. In Proceedings of the 30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, NC, USA, 30 October–1 November 1989; pp. 248–253. [CrossRef]
25. Mitzenmacher, M.; Vadhan, S. Why simple hash functions work: Exploiting the entropy in a data stream. In *Proceedings of the Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2008; pp. 746–755.
26. Bennett, C.; Brassard, G.; Crepeau, C.; Maurer, U. Generalized privacy amplification. *IEEE Trans. Inf. Theory* **1995**, *41*, 1915–1923. [CrossRef]
27. Barak, B.; Dodis, Y.; Krawczyk, H.; Pereira, O.; Pietrzak, K.; Standaert, F.X.; Yu, Y. Leftover Hash Lemma, Revisited. In Proceedings of the Advances in Cryptology—CRYPTO 2011—31st Annual Cryptology Conference, Santa Barbara, CA, USA, 14–18 August 2011; pp. 1–20. [CrossRef]
28. Dodis, Y.; Yu, Y. Overcoming Weak Expectations. In Proceedings of the Theory of Cryptography—10th Theory of Cryptography Conference (TCC 2013), Tokyo, Japan, 3–6 March 2013; pp. 1–22. [CrossRef]
29. Xu, D.; Erdogmuns, D. Renyi's Entropy, Divergence and Their Nonparametric Estimators. In *Information Theoretic Learning*; Information Science and Statistics; Springer: New York, NY, USA, 2010; pp. 47–102. [CrossRef]
30. Cachin, C. Smooth entropy and Rényi entropy. In *Proceedings of the 16th Annual International Conference on Theory and Application of cRyptographic Techniques (EUROCRYPT'97)*; Springer: Berlin/Heidelberg, Germany, 1997; pp. 193–208. [CrossRef]
31. Renner, R.; Wolf, S. Smooth Renyi entropy and applications. In Proceedings of the 2004 IEEE International Symposium on Information Theory (ISIT 2004), Chicago Downtown Marriott, Chicago, IL, USA, 27 June–2 July2004; p. 233. [CrossRef]
32. Acharya, J.; Orlitsky, A.; Suresh, A.T.; Tyagi, H. Estimating Renyi Entropy of Discrete Distributions. *IEEE Trans. Inf. Theory* **2017**, *63*, 38–56. [CrossRef]
33. Lake, D.E. Nonparametric Entropy Estimation Using Kernel Densities. In *Methods in Enzymology*; Elsevier: Amsterdam, The Netherlands, 2009; Volume 467, pp. 531–546. [CrossRef]
34. Acharya, J.; Orlitsky, A.; Suresh, A.T.; Tyagi, H. The Complexity of Estimating Rényi Entropy. In Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2015), San Diego, CA, USA, 4–6 January 2015; pp. 1855–1869. [CrossRef]
35. Obremski, M.; Skorski, M. Renyi Entropy Estimation Revisited. In *Approximation, Randomization, and Combinatorial Optimization, Algorithms and Techniques (APPROX/RANDOM 2017)*; Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany, 2017, Volume 81 Leibniz International Proceedings in Informatics (LIPIcs), pp. 20:1–20:15. [CrossRef]
36. Schaub, A.; Rioul, O.; Boutros, J.J. Entropy Estimation of Physically Unclonable Functions via Chow Parameters. In Proceedings of the 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 24–27 September 2019; pp. 698–704. [CrossRef]
37. Diakonikolas, I.; Gouleakis, T.; Peebles, J.; Price, E. Collision-Based Testers are Optimal for Uniformity and Closeness. *Chic. J. Theor. Comput. Sci.* **2019**, *25*, 1–21. [CrossRef]

38. Hoeffding, W. A Class of Statistics with Asymptotically Normal Distribution. *Ann. Math. Stat.* **1948**, *19*, 293–325. [CrossRef]
39. Ferguson, T.S. U-Statistics. In *Lecture Notes for Statistics*; University of California-Los Angeles: Los Angeles, CA, USA, 2005.
40. Lugosi, G.; Mendelson, S. Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey. *Found. Comput. Math.* **2019**, *19*, 1145–1190. [CrossRef]
41. Skorski, M. Machine Learning Examples. 2022. Available online: https://github.com/maciejskorski/ml_examples/blob/master/RenyiEntropyEstimation.ipynb (accessed on 2 November 2022).
42. Niemiro, W.; Pokarowski, P. Fixed precision MCMC estimation by median of products of averages. *J. Appl. Probab.* **2009**, *46*, 309–329. [CrossRef]
43. Lehmann, E.L.; Scheffé, H. Completeness, Similar Regions, and Unbiased Estimation: Part II. *Sankhyā Indian J. Stat. (1933–1960)* **1955**, *15*, 219–236.
44. Lehmann, E.L.; Scheffé, H. Completeness, Similar Regions, and Unbiased Estimation: Part I. *Sankhyā Indian J. Stat. (1933–1960)* **1950**, *10*, 305–340.
45. Goldreich, O.; Ron, D. On testing expansion in bounded-degree graphs. In *Proceedings of the Electronic Colloquium on Computational Complexity (ECCC)*; Springer: Berlin/Heidelberg, Germany, 2000; Volume 20.
46. Paninski, L. A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data. *IEEE Trans. Inf. Theory* **2008**, *54*, 4750–4755. [CrossRef]
47. Konca, S.; Idris, M.; Gunawan, H. p-Summable Sequence Spaces with Inner Products. *Bitlis Eren Univ. J. Sci. Technol.* **2015**, *5*. [CrossRef]
48. Fortuin, C.M.; Ginibre, J.; Kasteleyn, P.W. Correlation Inequalities on Some Partially Ordered Sets. *Commun. Math. Phys.* **1971**, *22*, 89–103. [CrossRef]