THE QUERY COMPLEXITY OF ESTIMATING ENTROPY

by

Jafar Jafarov

B.S., Computer Engineering, Boğaziçi University, 2014

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University
2016

THE QUERY COMPLEXITY OF ESTIMATING ENTROPY

APPROVED BY:

Prof. Ahmet Celal Cem Say . . . .
(Thesis Supervisor)

Assoc. Prof. Ali Taylan Cemgil . . . .

Assoc. Prof. Wolfgang Hörmann . . . .

DATE OF APPROVAL: 31.03.2016

To My Parents...

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor A. C. Cem Say for introducing me to the area of theoretical computer science. It was his simple yet deep and poetic introduction that inspired me to pursue science. His endless support and guidance was truly invaluable.

I am obliged to Ryan O'Donnell not only for proposing the topic or contributing greatly to this thesis but also for being an inspiration both as a person and as a scientist.

I am grateful to Ali Taylan Cemgil and Wolfgang Hörmann, my thesis committee members, for devoting their time and energy to this work.

I would like to express my gratitude to Jayadev Acharya for his helpful answers to my questions.

But above all, I would like to thank my family for their continual support and my friends for making my years at Boğaziçi University so extraordinary.

ABSTRACT

THE QUERY COMPLEXITY OF ESTIMATING ENTROPY

We investigate the query complexity of additively estimating entropy of a discrete probability distribution in two settings. Let $\boldsymbol{p}$ be an unknown probability distribution on $[n] := \{1, 2, \ldots n\}$, and define two kinds of queries: A SAMP query takes no input and returns $x \in [n]$ with probability $\boldsymbol{p}[x]$; a PMF query takes as input $x \in [n]$ and returns the value $\boldsymbol{p}[x]$. In the SAMP model of query complexity, the only allowed interaction with $\boldsymbol{p}$ is via SAMP queries. In the SAMP+PMF model, both SAMP and PMF queries are utilized to interact with $\boldsymbol{p}$.

In particular, we consider the task of estimating the entropy of $\boldsymbol{p}$ to within $\pm\Delta$ (with high probability). For the usual Shannon entropy $H(\boldsymbol{p})$, we review the matching upper and lower bounds established by Valiant and Valiant in the SAMP model, and describe the algorithm constructed by Canonne and Rubinfeld in the SAMP+PMF model. For the Rényi entropy $H_\alpha(\boldsymbol{p})$, we analyze three different matching upper and lower bound pairs introduced by Acharya *et al.* in the SAMP model.

We show that $\Omega(\log^2 n/\Delta^2)$ queries are necessary to estimate the Shannon entropy $H(\boldsymbol{p})$ in the SAMP+PMF model, matching a recent upper bound of Canonne and Rubinfeld. In addition, we prove that $\Theta\left(n^{1-1/\alpha}\right)$ queries are necessary and sufficient to estimate the Rényi entropy $H_\alpha(\boldsymbol{p})$ in the SAMP+PMF model, where $\alpha > 1$. This complements recent work of Acharya *et al.* in the SAMP model that showed $O(n^{1-1/\alpha})$ queries suffice when α is an integer, but roughly n queries are necessary when α is a noninteger. All of our lower bounds extend to the SAMP+CDF model, where SAMP and CDF queries (given x, return $\sum_{y \le x} \boldsymbol{p}[y]$) are allowed. We give a matching lower bound on estimating the support size (the number of domain elements with nonzero probability) of an unknown distribution $\boldsymbol{p}$ in the SAMP+CDF model. Lastly, we present an upper bound on additively estimating Tsallis entropy in the SAMP+PMF model.

ÖZET

ENTROPİ KESTİRİMİNİN SORGU KARMAŞIKLIĞI

Bu çalışmada, ayrık olasılık dağılımının entropisinin toplanır hata payıyla kestirimi iki farklı kurguda irdelenmektedir. Buna göre bilinmeyen bir olasılık dağılımı $\boldsymbol{p}$'ye erişim iki farklı sorgu türüyle sağlanmaktadır. Herhangi bir girdisi olmayan SAMP sorgusu $\boldsymbol{p}[x]$ olasılığıyla $x \in [n]$ döndürmektedir. Girdi olarak $x \in [n]$ alan PMF sorgusunun ise çıktısı $\boldsymbol{p}[x]$'dir. SAMP modeli ismini verdiğimiz ilk kurguda $\boldsymbol{p}$ ile sadece SAMP sorgusu vasıtasıyla iletişim sağlanmaktadır. SAMP+PMF modeli olarak adlandırdığımız ikinci kurgudaysa hem SAMP hem de PMF sorguları kullanılabilmektedir.

Daha kesin bir ifadeyle, bu çalışmanın odak noktası olasılık dağılımı $\boldsymbol{p}$'nin entropisinin yüksek ihtimalle $\pm\Delta$ toplanır hata payıyla kestirimi problemidir. Shannon entropisi $H(\boldsymbol{p})$'nin kestirimini incelediğimiz bölümde Valiant ve Valiant'ın SAMP modelinde göstermiş olduğu eşleşen alt ve üst sınırları ve Canonne ve Rubinfeld'in SAMP+PMF modelinde inşa etmiş olduğu algoritmayı tasvir ediyoruz. Rényi entropisi $H_\alpha(\boldsymbol{p})$'yi incelediğimiz bölümdeyse Acharya ve diğerleri tarafından sunulan üç farklı eşleşen alt ve üst sınır çiftini analiz ediyoruz.

Kendi katkımız olarak, önce SAMP+PMF modelinde Shannon entropisi $H(\boldsymbol{p})$'nin kestirimi probleminin $\Omega(\frac{\log^2 n}{\Delta^2})$ sayıda sorgu gerektirdiğini kanıtlayarak Canonne ve Rubinfeld'in sunduğu üst sınırın optimal olduğunu gösteriyoruz. İkinci olarak, yine SAMP+PMF modelinde Rényi entropisi $H_\alpha(\boldsymbol{p})$'yi $\alpha > 1$ değerlerinde kestirebilmek için $\Theta\left(n^{1-1/\alpha}\right)$ sayıda sorgunun gerekli ve yeterli olduğunu ispatlıyoruz. Böylelikle Acharya ve diğerleri tarafından yakın zamanda elde edilen, SAMP modelinde Rényi entropisi $H_\alpha(\boldsymbol{p})$'yi $\alpha > 1$ tamsayı değerlerinde kestirebilmek için $O\left(n^{1-1/\alpha}\right)$ sayıda sorgunun yeterli olduğu fakat $\alpha > 1$ tamsayı olmayan değerlerinde kestirebilmek için kabaca n sayıda sorgunun gerekli olduğu yönündeki sonuçları tamamlamış oluyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

$\mathrm{Bin}\left(\cdot,\cdot\right)$	Binomial distribution
$\mathbf{Cov}\left[\cdot\right]$	Covariance operator
$\exp(\cdot)$	Exponential operator
$\mathbf{E}\left[\cdot\right]$	Expected value operator
$\mathcal{F}_{\mathcal{X}}$	Fingerprint of a sample set $\mathcal{X}$
h	Histogram of a distribution $\boldsymbol{p}$
$H\left(\boldsymbol{p}\right)$	Shannon entropy of a distribution $\boldsymbol{p}$
$H_\alpha\left(\boldsymbol{p}\right)$	Rényi entropy of degree α of a distribution $\boldsymbol{p}$
$\mathbf{KL}\left(\cdot\right)$	Kullback-Leibler divergence
$\mathcal{M}_\alpha\left(\boldsymbol{p}\right)$	Moment of degree α of a distribution $\boldsymbol{p}$
$\widehat{\mathcal{M}^e_\alpha}$	Empirical estimator of $\mathcal{M}_\alpha\left(\boldsymbol{p}\right)$
$\widehat{\mathcal{M}^u_\alpha}$	Bias-corrected estimator of $\mathcal{M}_\alpha\left(\boldsymbol{p}\right)$
$[n]$	Set of natural numbers $\{1,\ldots,n\}$
N_i	Multiplicity of a domain element i
$\boldsymbol{p}$	Discrete probability distribution on domain $[n]$
$\mathrm{Pois}\left(\cdot\right)$	Poisson distribution
$\mathrm{supp}\left(\boldsymbol{p}\right)$	Support size of a distribution $\boldsymbol{p}$
$\mathcal{U}\left([n]\right)$	Uniform distribution on the domain $[n]$
$\mathbf{Var}\left[\cdot\right]$	Variance operator
$\mathcal{X}$	Set of m independent samples $\{X_1,\ldots\ X_m\}$
δ	Error probability
Δ	Additive accuracy
$\mathbb{1}_E$	Indicator function of an event E

LIST OF ACRONYMS/ABBREVIATIONS

CLT	Central Limit Theorem
CDF	Cumulative Distribution Function
LP	Linear Program
KL	Kullback-Leibler
PMF	Probability Mass Function
SAMP	Sampling

1. INTRODUCTION

The question of what to infer about an unknown probability distribution $\boldsymbol{p}$ given samples from it is fundamental to the field of statistics and has been researched for decades. However, the practicality of traditional techniques has been shaken due to the rapidly growing size of data in scientific study, a phenomenon designated as big data. In the absence of simplifying assumptions about a probability distribution such as being of a specific type or possessing certain smoothness properties, the number of samples utilized by such techniques grows linearly in the size of the domain of a distribution which is huge in the realm of big data. Thus, the task of constructing algorithms with sublinear sample complexity has become all-important, and the aforementioned question has been recently investigated within the theoretical computer science framework of *property testing*. In this framework, as its name implies, a certain characteristic of a distribution is put under the microscope. In addition, the only assumption made about $\boldsymbol{p}$ is that it is a discrete probability distribution on a finite domain $[n] := \{1, 2, \ldots, n\}$ where $n \in \mathbb{N}$. For a detailed exploration of the field, see the surveys by Rubinfeld [1] and Canonne [2].

One of the most significant characteristics of a probability distribution is its Shannon entropy, $H(\boldsymbol{p}) = -\sum_{i=1}^{n} \boldsymbol{p}[i] \log \boldsymbol{p}[i]$,[1] which represents the "amount of randomness" a distribution possesses. The first focal point of this work is estimating Shannon entropy to within additive error Δ with probability at least $1 - \delta$. (In a typical scenario $\Delta = 1$ and $\delta = 1/3$.) The reason we limit ourselves to additive rather than multiplicative estimation is that it is directly related to the estimation of mutual information. That is, if $\boldsymbol{p}$ is a joint probability distribution of two random variables X, Y, then additively estimating mutual information $I(X, Y) = H(X) + H(Y) - H(X, Y)$ is realized via additively estimating $H(\boldsymbol{p})$. For deeper analysis of Shannon entropy and mutual information see Paninski [3]. The second focal point of this work is estimating another popular type of entropy, Rényi entropy $H_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} \boldsymbol{p}[i]^\alpha \right)$, to within additive error Δ with probability at least $1 - \delta$, where $\alpha \in [0, 1) \cup (1, \infty)$. Both tasks

[1] In this work, log denotes $\log_2$.

are investigated in two different settings.

In the conventional setting, which we refer to as the SAMP model, the only allowed interaction with a probability distribution $\boldsymbol{p}$ is via independent samples. As recently shown in [3–6], $\Theta\left(\frac{n}{\log n}\right)$ samples are necessary and sufficient to estimate Shannon entropy to within a constant additive error with high probability. The case for estimating Rényi entropy is more complicated; three different results for three classes of α are obtained in [7], the most efficient one being for the class of integer $\alpha > 1$ with $\Theta\left(n^{1-1/\alpha}\right)$ sample complexity. These quantities are not always convenient, considering the aforementioned tendency in science and technology.

In the "unconventional" setting referred to as the SAMP+PMF model, aside from drawing independent samples as in the SAMP model, querying a probability mass function (PMF)[2] of an arbitrary domain element, that is, learning $\boldsymbol{p}[i]$ of an element $i \in [n]$ is allowed. This extended version of the SAMP model, called the "Generation+Evaluation" model in [8] and the "combined model" in [9], is introduced to overcome the difficulty described above. The results achieved in [10] imply that estimating Shannon entropy is possible with polylog(n) SAMP+PMF queries, exponentially better than the $\Omega\left(\frac{n}{\log n}\right)$ queries in the SAMP model.

Although described as unconventional, the SAMP+PMF model becomes practical in many applications. For instance, the number of occurrences, and therefore the probability of a certain element in a sorted database can be calculated via at most logarithmically many interactions with the database. For a concrete example, consider the Google n-gram database in which the frequency of each n-gram is published, and a random n-gram is easily obtained from the underlying text corpus. Another motivation for the SAMP+PMF model stems from its strong relation with the *streaming* model [11], where entropy estimation has been thoroughly studied [12–17]. To exhibit the relation between the two, note that any q-query estimation algorithm in the SAMP+PMF model can be converted to a $q \cdot \text{polylog}(n)$-space streaming algorithm with one or two passes

[2] In this work, PMF, CDF and SAMP are abbreviations for probability mass function, cumulative distribution function and sampling, respectively.

(details of the conversion depend on the model for how the items in the stream are ordered). For more motivation and results for the SAMP+PMF model, see Canonne and Rubinfeld [10].

Our main contribution [18] is to establish a lower bound matching the upper bound obtained in [10] on additively estimating Shannon entropy, $\Omega\left(\log^2 n\right)$. In addition, we found upper and lower bounds matching in their dependence on n for additive estimation of Rényi entropy when $\alpha > 1$, $\Theta\left(n^{1-1/\alpha}\right)$.

1.1. Organization

In Chapter 2, we focus on the task of additively estimating entropy in the SAMP model. Section 2.1 is devoted to rigorously analyzing optimal upper and lower bounds achieved in three successive works [19–21]. Section 2.2 explores three different matching upper and lower bound pairs on additive estimation of Rényi entropy obtained in [7].

In Chapter 3, we concentrate on additively approximating the Shannon entropy in the SAMP+PMF model. We describe the exponentially better algorithm constructed by Canonne and Rubinfeld [10]. In addition, we introduce the SAMP+CDF model, which is an extension of the SAMP+PMF model.

In Chapter 4, we demonstrate our contribution to the problem. Section 4.1 includes a comparison between our results and the prior work. In Section 4.2, we build an optimal lower bound on additively estimating Shannon entropy, and in Section 4.3, we present upper and lower bounds on additive estimation of Rényi entropy in the SAMP+PMF and SAMP+CDF models. In Section 4.4, we establish a lower bound on estimating the support size of a probability distribution in the SAMP+CDF model. Section 4.5 is devoted to constructing an algorithm for additively estimating Tsallis entropy in the SAMP+PMF model.

Finally, we state some open questions in the conclusion and give the inequalities used throughout this work in the appendix.

2. SAMP MODEL

We start with the formal definition of the conventional model.

Definition 2.1. *(sampling-only model) Let $\boldsymbol{p}$ be a probability distribution on $[n]$ and SAMP denote a type of query which takes no input and returns $i \in [n]$ with probability $\boldsymbol{p}[i]$ independently of all previous calls. The SAMP model is a model of query complexity in which the only allowed interaction with $\boldsymbol{p}$ is via SAMP queries.*

Although there are different metrics to measure the distance between two probability distributions $\boldsymbol{p_1}, \boldsymbol{p_2}$, the most commonly used metric is *total variation distance*, denoted by $\mathrm{d_{TV}}$ and defined as

$$\mathrm{d_{TV}}\left(\boldsymbol{p_1}, \boldsymbol{p_2}\right) := \frac{1}{2}\|\boldsymbol{p_1} - \boldsymbol{p_2}\|_1 = \frac{1}{2}\sum_{i=1}^{n} \left|\boldsymbol{p_1}[i] - \boldsymbol{p_2}[i]\right|. \tag{2.1}$$

The following identity unveils the "mystery" behind the constant factor $\frac{1}{2}$.

$$\mathrm{d_{TV}}\left(\boldsymbol{p_1}, \boldsymbol{p_2}\right) = \max_{E \subseteq [n]} \left\{\boldsymbol{p_1}(E) - \boldsymbol{p_2}(E)\right\} \tag{2.2}$$

Proof. Let $A = \{i : \boldsymbol{p_1}[i] \geq \boldsymbol{p_2}[i]\}$. Then

$$\mathrm{d_{TV}}\left(\boldsymbol{p_1}, \boldsymbol{p_2}\right) = \frac{1}{2}\left(\boldsymbol{p_1}(A) - \boldsymbol{p_2}(A)\right) + \frac{1}{2}\left(\boldsymbol{p_2}\left(\overline{A}\right) - \boldsymbol{p_1}\left(\overline{A}\right)\right) = \boldsymbol{p_1}(A) - \boldsymbol{p_2}(A),$$

where $\overline{A} := [n] \setminus A$. The next step is to show that A is an event maximizing the right-hand side of Equation 2.2. If one adds another element $j \in [n]$ to A, the difference $\boldsymbol{p_1}(A) - \boldsymbol{p_2}(A)$ decreases, since by definition $\boldsymbol{p_1}[j] < \boldsymbol{p_2}[j]$. Similarly, removing an element j from A leads to a decrease in $\boldsymbol{p_1}(A) - \boldsymbol{p_2}(A)$, since $\boldsymbol{p_1}[j] \geq \boldsymbol{p_2}[j]$. □

2.1. Shannon Entropy

Shannon entropy, named after Claude E. Shannon [22], represents the expected information a probability distribution contains, thus, measures the randomness in a distribution and the compressibility of the data produced by that distribution. Shannon entropy is defined as

$$H\left(\boldsymbol{p}\right) = -\sum_{i=1}^{n} \boldsymbol{p}\left[i\right] \log \boldsymbol{p}\left[i\right]. \tag{2.3}$$

By convention, the quantity $\boldsymbol{p}\left[i\right] \log \boldsymbol{p}\left[i\right]$ is set to 0 in the case of $\boldsymbol{p}\left[i\right] = 0$ for some i which is consistent with the following: $\lim\limits_{x \to 0^+} x \log x = 0$. Note that

$$0 \leq H\left(\boldsymbol{p}\right) \leq \log n. \tag{2.4}$$

The left-hand side of Inequality 2.4 is trivial, since $-\boldsymbol{p}\left[i\right] \log \boldsymbol{p}\left[i\right] \geq 0$ for all i. The right-hand side of Inequality 2.4 follows from the fact that $H\left(\boldsymbol{p}\right) = \mathbf{E}_{i \sim \boldsymbol{p}} \left[\log \frac{1}{\boldsymbol{p}[i]}\right]$ and $\log x$ is a concave function. By applying Jensen's inequality,[3]

$$H\left(\boldsymbol{p}\right) = \mathop{\mathbf{E}}_{i \sim \boldsymbol{p}} \left[\log \frac{1}{\boldsymbol{p}\left[i\right]}\right] \leq \log \left(\mathop{\mathbf{E}}_{i \sim \boldsymbol{p}} \left[\frac{1}{\boldsymbol{p}\left[i\right]}\right]\right) = \log n. \tag{2.5}$$

Shannon entropy has many applications such as measuring genetic diversity [23], quantifying neural activity [3], and detecting network anomalies [13].

This work concentrates on additive approximation to the entropy, though we also state the results regarding its multiplicative counterpart. Batu, Dasgupta, Kumar and Rubinfeld [9] construct an algorithm approximating $H\left(\boldsymbol{p}\right)$ of a distribution $\boldsymbol{p}$ within a multiplicative factor of γ using $\tilde{O}\left(n^{(1+o(1))/\gamma^2}\right)$ samples given $H\left(\boldsymbol{p}\right) = \Omega\left(\gamma\right)$ for any $\gamma > 1$. They also show that no algorithm exists which $\gamma-$ approximates the entropy of every distribution. Furthermore, Valiant [5] proves that $\Omega\left(n^{1/\gamma^2}\right)$ samples are necessary for the task given $H\left(\boldsymbol{p}\right) = \Omega\left(\frac{\log n}{\gamma^2}\right)$.

[3] For the definitions of the inequalities used throughout this work, see Appendix A.

Initially, we introduce a trivial upper bound on additively estimating Shannon entropy.

Fact 2.2. *There exists an algorithm estimating (with high probability) the Shannon entropy to within arbitrarily small constant* Δ *using* $O\left(\frac{n \log^2 n}{\Delta^2}\right)$ *samples.*

Proof. The overall idea is to build a distribution $\boldsymbol{p}'$ which is close to $\boldsymbol{p}$ in total variation distance and output $H(\boldsymbol{p}')$ as an approximation of $H(\boldsymbol{p})$. We utilize the "plug-in" distribution for $\boldsymbol{p}'$. Namely, let $X_1, \ldots, X_m$ be m independent samples drawn from $\boldsymbol{p}$ and define $\boldsymbol{p}'[i] = \frac{1}{m} \sum\limits_{j=1}^{m} \mathbb{1}_{\{X_j = i\}}$ for each $i \in [n]$ where $\mathbb{1}_E$ is the indicator function for an event E. Observe that $\boldsymbol{p}'[i]$ is an unbiased estimator of $\boldsymbol{p}[i]$ since $\mathbf{E}[\boldsymbol{p}'[i]] = \boldsymbol{p}[i]$ for each i. To show that $|H(\boldsymbol{p}') - H(\boldsymbol{p})| \leq \Delta$ we use the following fact:

Fact 2.3. ([24], Lemma 8) *Let* $\boldsymbol{p_1}, \boldsymbol{p_2}$ *be two arbitrary probability distributions such that* $\mathrm{d_{TV}}(\boldsymbol{p_1}, \boldsymbol{p_2}) \leq \frac{\Delta}{4 \log n}$, *then* $|H(\boldsymbol{p_1}) - H(\boldsymbol{p_2})| \leq \Delta$ *where* Δ *is an arbitrarily small constant.*

The final step is to prove that $\mathrm{d_{TV}}(\boldsymbol{p}', \boldsymbol{p}) \leq \frac{\Delta}{4 \log n}$ with high probability when $m = O\left(\frac{n \log^2 n}{\Delta^2}\right)$. Similarly, $\boldsymbol{p}'[E] := \frac{1}{m} \sum\limits_{j=1}^{m} \mathbb{1}_{\{X_j \in E\}}$ is an unbiased estimator of $\boldsymbol{p}[E]$ for any event $E \subseteq [n]$. Observe that $\Pr\left[\left|\boldsymbol{p}'[E] - \boldsymbol{p}[E]\right| > \frac{\Delta}{4 \log n}\right] \leq 2e^{-\frac{\Delta^2}{8 \log^2 n} m} = e^{-O(n)}$ for such m, which is easily derived via the Hoeffding bound. Then,

$$
\begin{aligned}
\Pr\left[\mathrm{d_{TV}}(\boldsymbol{p}', \boldsymbol{p}) > \frac{\Delta}{4 \log n}\right] &= \Pr\left[\max_{E \subseteq [n]} \{\boldsymbol{p}'(E) - \boldsymbol{p}(E)\} > \frac{\Delta}{4 \log n}\right] \\
&\leq \Pr\left[\bigcup_{E \subseteq [n]} \left\{\boldsymbol{p}'(E) - \boldsymbol{p}(E) > \frac{\Delta}{4 \log n}\right\}\right] \\
&\leq \sum_{E \subseteq [n]} \Pr\left[\boldsymbol{p}'(E) - \boldsymbol{p}(E) > \frac{\Delta}{4 \log n}\right] \\
&\leq 2^n \cdot e^{-O(n)} = o(1).
\end{aligned}
$$

□

The task of achieving an additive estimation of entropy is fulfilled in its entirety in three successive works [19], [20] and [21]. Valiant and Valiant establish matching upper and lower bounds on additively estimating a major class of symmetric properties (a property is symmetric if it is immune to any permutation of domain elements) including entropy. It is proven that $\Theta\left(\frac{n}{\log n}\right)$ samples are necessary and sufficient for additive estimation of entropy of a probability distribution with support size at most n. We use a rather technical narrative for two reasons. First, each work has an intricate structure requiring an attentive analysis. Second, the techniques deployed in the process utilize a wide range of mathematical notions that may be of independent interest.

2.1.1. Upper Bound I

In this part, we introduce the algorithm constructed in [20] estimating entropy up to an arbitrarily small constant using $O\left(\frac{n}{\log n}\right)$ independent samples. As the title of the paper (*Estimating the unseen: A sublinear-sample canonical estimator of distributions*) suggests, Valiant and Valiant employ a canonical approach to delicately approximate an unobserved portion of a probability distribution rather than directly estimating entropy. In other words, based on the samples drawn from an unknown probability distribution $\boldsymbol{p}$, the estimator builds a probability distribution $\boldsymbol{p}'$ such that with high probability the two are "close", and returns the entropy of $\boldsymbol{p}'$ as an approximation of the entropy of $\boldsymbol{p}$. The success of obtaining sublinear-sample complexity is due to exploiting features of symmetry and using a different distance metric to better capture the "closeness" between $\boldsymbol{p}$ and $\boldsymbol{p}'$. Before going into the details, we introduce some key definitions.

Definition 2.4. *A property of a distribution is a function $\pi : \boldsymbol{p}^n \to \mathbb{R}$, where $\boldsymbol{p}^n$ is the set of distributions on domain $[n]$. A property π is called a* ***symmetric property*** *if for all distributions $\boldsymbol{p}$, and all permutations σ, $\pi(\boldsymbol{p}) = \pi(\boldsymbol{p} \circ \sigma)$.*

Note that entropy is a symmetric property.

Definition 2.5. *Given a sequence of samples $\mathcal{X} = \{X_1, \ldots, X_m\}$, let the associated* ***fingerprint****, denoted by $\mathcal{F}_{\mathcal{X}}$, be the vector whose i^{th} component, $\mathcal{F}_{\mathcal{X}}(i)$ is the number of domain elements that occur exactly $i \geq 1$ times in sample $\mathcal{X}$.*

Intuitively, the fingerprint of a sample should hold all necessary information about a sample for the task of estimating a symmetric property. We formalize this intuition via the following fact.

Fact 2.6. ([9], Lemma 8) *For any algorithm $\mathcal{A}$ that approximates the entropy of a distribution to within additive Δ from samples, there exists an algorithm $\mathcal{A}'$ which gets as input only the fingerprint of the generated sample and has the error probability upper bounded by that of $\mathcal{A}$.*

Proof. Let $\boldsymbol{p}$ be an unknown distribution and $\mathcal{F}_{\mathcal{X}}$ denote the fingerprint of the set of samples $\mathcal{X} = \{X_1, \ldots, X_m\}$ drawn from $\boldsymbol{p}$. Algorithm $\mathcal{A}'$ is constructed as follows:

- Choose $\mathcal{F}_{\mathcal{X}}(i)$ elements at random from $[n]$ without replacement for each i,[4]
- Build $\mathcal{X}'$ so that an element chosen in step i occurs exactly i times in $\mathcal{X}'$,
- Output the value that $\mathcal{A}$ outputs on $\mathcal{X}'$.

The next step is to prove the correctness of $\mathcal{A}'$. Let π be a permutation on $[n]$ and define a permuted distribution $\pi(\boldsymbol{p})$ such that $\pi(\boldsymbol{p})[i] = \boldsymbol{p}[\pi(i)]$. Let $\pi(\mathcal{X})$ be a set of samples by relabeling the members of $\mathcal{X}$ according to π. Observe that the set $\mathcal{X}'$ generated by $\mathcal{A}'$ is $\pi(\mathcal{X})$ for some random permutation π. Lastly, let $\mathcal{A}(\mathcal{X})$ denote the output of $\mathcal{A}$ on the sample set $\mathcal{X}$. Then,

$$\begin{aligned}
&\Pr\left[\mathcal{A}' \text{ estimates } H(\boldsymbol{p}) \text{ to within } \Delta\right] \\
&= \sum_{\mathcal{X}} \Pr\left[\boldsymbol{p} \text{ generates } \mathcal{X}\right] \cdot \mathop{\mathbf{E}}_{\pi}\left[\Pr\left[\mathcal{A}(\pi(\mathcal{X})) \text{ estimates } H(\boldsymbol{p}) \text{ to within } \Delta\right]\right] \\
&= \mathop{\mathbf{E}}_{\pi}\left[\sum_{\mathcal{X}} \Pr\left[\boldsymbol{p} \text{ generates } \mathcal{X}\right] \cdot \Pr\left[\mathcal{A}(\pi(\mathcal{X})) \text{ estimates } H(\boldsymbol{p}) \text{ to within } \Delta\right]\right] \\
&= \mathop{\mathbf{E}}_{\pi}\left[\sum_{\mathcal{X}} \Pr\left[\pi(\boldsymbol{p}) \text{ generates } \pi(\mathcal{X})\right] \cdot \Pr\left[\mathcal{A}(\pi(\mathcal{X})) \text{ estimates } H(\boldsymbol{p}) \text{ to within } \Delta\right]\right] \\
&= \mathop{\mathbf{E}}_{\pi}\left[\Pr\left[\mathcal{A} \text{ estimates } H(\pi(\boldsymbol{p})) \text{ to within } \Delta\right]\right] \\
&\geq \min_{\pi} \Pr\left[\mathcal{A} \text{ estimates } H(\pi(\boldsymbol{p})) \text{ to within } \Delta\right],
\end{aligned}$$

[4] Note that $n - \|\mathcal{F}_{\mathcal{X}}\|_1$ is the number of elements not seen in the sample set $\mathcal{X}$.

which is the correctness probability of $\mathcal{A}$. □

Similarly we define a histogram of the distribution which categorizes the domain elements according to their probability values.

Definition 2.7. *The* ***histogram*** *of a distribution* $\boldsymbol{p}$ *is a mapping* $h : (0, 1] \to \mathbb{Z}$*, where* $h(x) = |\{i : \boldsymbol{p}[i] = x\}|$*. Additionally, generalized histograms are allowed which do not necessarily take integral values.*

Observe that, a symmetric property is a function of the histogram of a distribution. For instance, Shannon entropy can be written as

$$H(\boldsymbol{p}) = -\sum_{i=1}^{n} \boldsymbol{p}[i] \log \boldsymbol{p}[i] = \sum_{x:h(x)\neq 0} h(x)\, x \log \frac{1}{x}. \tag{2.6}$$

We now define a new distance metric to obtain a better measure for proximity between distributions.

Definition 2.8. *For two histograms (or generalized histograms)* h_1, h_2*, let the* ***relative earthmover distance*** *between them,* $R(h_1, h_2)$*, be the minimum over all schemes of moving the probability mass of the first histogram to yield the second histogram, of the cost of moving that mass, where the per-unit cost of moving mass from probability* x *to* y *is* $|\log(x/y)|$*.*

Distributions which are close to each other according to this new metric have similar entropies.

Definition 2.9. *A symmetric property* π *is* (Δ, ϑ)*-continuous if for all distributions* $\boldsymbol{p_1}, \boldsymbol{p_2}$ *with respective histograms* h_1, h_2 *satisfying* $R(h_1, h_2) \leq \vartheta$ *it follows that*

$$|\pi(\boldsymbol{p_1}) - \pi(\boldsymbol{p_2})| \leq \Delta. \tag{2.7}$$

Fact 2.10. ([20], Fact 9) *For a probability distribution* $\boldsymbol{p}$*, and* $\Delta > 0$ *the Shannon entropy,* $H(\boldsymbol{p}) = -\sum\limits_{i=1}^{n} \boldsymbol{p}[i] \log \boldsymbol{p}[i]$ *is* (Δ, Δ)*-continuous, with respect to the relative earthmover distance.*

We describe a well-known sampling technique known as *Poisson sampling.* Recall that in the standard approach of estimating a certain $\boldsymbol{p}[i]$, one needs to draw m independent samples $X_1, \ldots, X_m$ from $\boldsymbol{p}$ and calculate the multiplicity of a domain element i, that is, the number of occurrences of i among m samples. Observe that the multiplicities of any two elements are not independent, complicating the overall analysis: for a start, the sum of the multiplicities of all domain elements must be equal to m. To overcome this difficulty, instead of drawing exactly m independent samples from $\boldsymbol{p}$ we draw $M \sim \text{Pois}(m)$ samples, where $\text{Pois}(m)$ is the Poisson distribution with parameter m. Let $X_1, \ldots, X_M$ be independent samples drawn from $\boldsymbol{p}$, then the multiplicity of a domain element i is defined as

$$N_i = |\{1 \leq j \leq M : X_j = i\}|. \tag{2.8}$$

Fact 2.11. *The multiplicities $\{N_i\}$ are independent random variables and distributed as Pois $(m \cdot \boldsymbol{p}[i])$.*

Proof. Let $\text{Bin}(y, z)$ be a Binomial distribution with parameters $y \in \text{N}$ and $z \in [0, 1]$. Then

$$\begin{aligned}
\Pr[N_i = j] &= \sum_{M=0}^{\infty} \left(\Pr[\text{Pois}(m) = M] \cdot \Pr[\text{Bin}(M, \boldsymbol{p}[i]) = j]\right) \\
&= \sum_{M=j}^{\infty} \left(\Pr[\text{Pois}(m) = M] \cdot \Pr[\text{Bin}(M, \boldsymbol{p}[i]) = j]\right) \\
&= \sum_{M=j}^{\infty} \left(\frac{e^{-m} m^M}{M!} \cdot \frac{M!}{j!\,(M-j)!} \cdot \boldsymbol{p}[i]^j \,(1 - \boldsymbol{p}[i])^{M-j}\right) \\
&= \frac{e^{-m} \boldsymbol{p}[i]^j}{j!} \sum_{M'=0}^{\infty} \left(\frac{m^{M'+j}}{M'!} \cdot (1 - \boldsymbol{p}[i])^{M'}\right) \\
&= \frac{e^{-m} \cdot (m\boldsymbol{p}[i])^j}{j!} \sum_{M'=0}^{\infty} \left(\frac{(m\,(1 - \boldsymbol{p}[i]))^{M'}}{M'!}\right) \\
&= \frac{e^{-m} \cdot (m\boldsymbol{p}[i])^j \cdot e^{m - m\boldsymbol{p}[i]}}{j!} \\
&= \frac{e^{-m\boldsymbol{p}[i]} \cdot (m\boldsymbol{p}[i])^j}{j!} = \text{Pois}(m\boldsymbol{p}[i], j).
\end{aligned}$$

Now we show that the N_i's are independent.

$$\begin{aligned}
\Pr\left[N_1 = M_1 \;\&\; \cdots \;\&\; N_n = M_n\right] &= \Pr\left[M\right] \cdot \Pr\left[N_1 = M_1 \;\&\; \cdots \;\&\; N_n = M_n \middle|\; M\;\right] \\
&= \frac{e^{-m} m^M}{M!} \cdot \frac{M!}{M_1! \cdots M_n!} \cdot \boldsymbol{p}\left[1\right]^{M_1} \cdots \boldsymbol{p}\left[n\right]^{M_n} \\
&= \frac{e^{-m\boldsymbol{p}[1]} \left(m\boldsymbol{p}\left[1\right]\right)^{M_1}}{M_1!} \cdots \frac{e^{-m\boldsymbol{p}[n]} \left(m\boldsymbol{p}\left[n\right]\right)^{M_n}}{M_n!} \\
&= \Pr\left[N_1 = M_1\right] \cdots \Pr\left[N_n = M_n\right]. \qquad \square
\end{aligned}$$

Observe that N_i/m is an unbiased estimator for $\boldsymbol{p}\left[i\right]$ because $\mathbf{E}[\frac{N_i}{m}] = \dfrac{\mathbf{E}[N_i]}{m} = \boldsymbol{p}\left[i\right]$. Consider the distribution of the j^{th} entry of a $\text{Pois}\left(m\right)$-sample fingerprint,

$$\mathcal{F}\left(j\right) = \sum_{i=1}^{n} \mathbb{1}\left\{N_i = j\right\} \;\Rightarrow\; \mathbf{E}\left[\mathcal{F}\left(j\right)\right] = \sum_{i=1}^{n} \text{Pois}\left(m\boldsymbol{p}\left[i\right], j\right) = \sum_{x:h(x)\neq 0} h\left(x\right) \text{Pois}\left(mx, j\right). \quad (2.9)$$

The direct consequence of N_i's being independent is that for each j, $\mathcal{F}\left(j\right)$ is closely concentrated around its expectation, having an easy proof by a direct application of the Chernoff bound. The other advantage of Poisson sampling is that its sample complexity is comparable to the sample complexity of usual sampling, because Poisson distribution also has a concentration around its expectation.

Proposition 2.12. ([20], Proposition 21) *Given $m > 30$, and any set of fingerprints A, let $\overline{A}$ be the set of fingerprints that can be obtained by adding or removing at most $m^{0.6}$ samples from some fingerprint in set A. Let $\mathcal{F}$ denote a random m-sample fingerprint, and let $\mathcal{F}'$ denote a fingerprint obtained from choosing $M \sim \text{Pois}\left(m\right)$, random samples. Then*

$$\Pr\left[\mathcal{F} \in A\right] \leq \Pr\left[\mathcal{F}' \in \overline{A}\right] + e^{-m^{0.1}/2}.$$

The construction of the estimator starts with building a linear program that focuses on the low probability portion of a distribution. Based on a fingerprint $\mathcal{F}$ of an unknown histogram h, the objective is to derive a histogram h' such that for each domain element i and for a fingerprint $\mathcal{F}_{\mathcal{X}}(i)$ obtained from a sample $\mathcal{X}$ of h', $\mathbf{E}\left[\mathcal{F}_{\mathcal{X}}(i)\right] \approx \mathcal{F}(i)$. However, for the high probability portion of a distribution, that is, for elements with probability at least $m^{-1+\alpha}$ for some small constant $\alpha \in (0,1)$, the histogram is set as $h'\left(\frac{j}{m}\right) = \mathcal{F}(j)$.

Definition 2.13 (The Linear Program LP). *Given an m-sample fingerprint $\mathcal{F}$ and $\alpha = 1/50, c \in [1,2]$, bounds $A := cm^{-1+\alpha}, B := 4m^{-1+0.6\alpha}$, and a real number $\gamma := m^{-3/2}$, the linear program consists of variables $v_x \geq 0$ for all $x \leq A + B/2$ in the set $X := \{\gamma, 2^2\gamma, 3^2\gamma, \ldots, A + B/2\}$, subject to the following condition:*

1. $\sum\limits_{x \in X : x \geq A} x v_x \leq 16 m^{-0.4\alpha}$
2. $\sum\limits_{x \in X} x v_x + \sum\limits_{j \geq m(A+B)} \frac{j}{m} \mathcal{F}_j = 1$
3. *For all integers $i \leq m\,(A + B/4)$,*

$$\sum_{x \in X} v_x \operatorname{Pois}(mx, i) \in \left[\mathcal{F}(i) - 4m^{0.6+\alpha}, \mathcal{F}(i) + 4m^{0.6+\alpha}\right].$$

The set X is chosen carefully to adjust the time complexity of LP to be linear in the number of samples. The first constraint is to guarantee that the probability mass residing in the neighborhood of the threshold probability, $A \approx m^{-1+\alpha}$, is small. The second constraint is to guarantee that the sum of the total probability mass of rarely occurring elements and the total probability mass of frequently occurring elements is 1. The third constraint is to guarantee that for rarely occurring domain elements the expectation of a fingerprint distribution $\mathcal{F}'$ of a histogram h' is close to a fingerprint $\mathcal{F}$ of a histogram h.

Observe that a solution $\boldsymbol{v} = \{v_x\}$ does not necessarily yield a proper histogram h' since v_x's can be nonintegers. The following definition is to construct a proper histogram which is referred as the histogram associated to a solution $\boldsymbol{v}$.

Definition 2.14. *Let $X := \{\gamma, 2^2\gamma, 3^2\gamma, \ldots, A + B/2\}$ be the set of probabilities for which LP solves. Given a $m-$fingerprint $\mathcal{F}$ and a solution $\boldsymbol{v}$ to the associated LP, the corresponding histogram $h^{\boldsymbol{v}}$ is derived from $\boldsymbol{v}$ according to the following process.*

1. *set* $h^{\boldsymbol{v}}(*) = 0$.
2. *for all* $x \in X$ *let* $h^{\boldsymbol{v}}(x) = v_x$.
3. *for all integers* $j \geq m(A+B)$, *let* $h^{\boldsymbol{v}}\left(\frac{j}{m}\right) = \mathcal{F}(j)$.
4. *for all* x *such that* $h^{\boldsymbol{v}}(x) \neq 0$, *set* $h^{\boldsymbol{v}}((1+\epsilon)x) = \lfloor h^{\boldsymbol{v}}(x) \rfloor$ *where* $\epsilon = \frac{\sum\limits_{x \in X} x(v_x - \lfloor v_x \rfloor)}{1 - \sum\limits_{x \in X} x(v_x - \lfloor v_x \rfloor)}$.

The first and second steps assign $\boldsymbol{v}$ to $h^{\boldsymbol{v}}$. The third step makes the histogram $h^{\boldsymbol{v}}$ for frequently occurring elements agree with the $\mathcal{F}$ of a histogram h. The last step converts the histogram values to integers while compensating the resulting loss in total probability mass by renormalizing the distribution.

Lastly, we establish the connection between LP and the aforementioned $O\left(\frac{n}{\log n}\right)$ sample complexity.

Algorithm: ESTIMATOR I

- Fix $m = O\left(\frac{n}{\log n}\right)$. (For details see [20].)
- Draw $M \sim \text{Pois}(m)$ independent samples $X_1, \ldots, X_M$ from $\boldsymbol{p}$.
- Construct LP corresponding to the fingerprint $\mathcal{F}$ obtained from $X_1, \ldots, X_M$.
- Find a solution $\boldsymbol{v}$ to LP.
- Compute histogram $h^{\boldsymbol{v}}$ associated to solution $\boldsymbol{v}$, as defined in Definition 2.14.
- Output $\sum\limits_{x : h^{\boldsymbol{v}}(x) \neq 0} h^{\boldsymbol{v}}(x)\, x \log \frac{1}{x}$.

Figure 2.1. Canonical Estimator of Shannon Entropy.

Theorem 2.15. ([20], Theorem 2) *For a constant* $\varepsilon \in (0,1]$, *consider a sample consisting of* m *independent samples from a histogram* h *of support size at most* $\varepsilon m \log m$. *With probability at least* $1 - e^{-m^{0.04}}$, *LP has a solution and furthermore, for any solution to LP,* $\boldsymbol{v}$, *the histogram* $h^{\boldsymbol{v}}$ *associated to* $\boldsymbol{v}$ *as in Definition 2.14 satisfies*

$$R(h, h^{\boldsymbol{v}}) = O\left(\sqrt{\varepsilon} \cdot \max\{1, |\log \varepsilon|\}\right).$$

Theorem 2.15 combining with Fact 2.10 imply that entropy of the resulting histogram $h^{\boldsymbol{v}}$ is close to the entropy of an unknown histogram h. The proof of Theorem 2.15 consists of two parts; in the first part it is shown that with the claimed probability there exists a feasible point $\widehat{\boldsymbol{v}}$ such that the associated histogram $h^{\widehat{\boldsymbol{v}}}$ is close to h. We informally describe how the existence of a feasible point $\widehat{\boldsymbol{v}}$ is proven.

A feasible point $\widehat{\boldsymbol{v}}$ is manually constructed via discretizing the low probability portion of an unknown histogram h. For each probability $y \leq A + \frac{B}{2}$ let x_i, x_{i+1} be consecutive members of the set X such that $x_i \leq y \leq x_{i+1}$, then $h\left(y\right) > 0$ is distributed between $\widehat{\boldsymbol{v}}_{\boldsymbol{x_i}}$ and $\widehat{\boldsymbol{v}}_{\boldsymbol{x_{i+1}}}$ as follows:

All initially being equal to 0, set $\widehat{\boldsymbol{v}}_{\boldsymbol{x_i}} := \widehat{\boldsymbol{v}}_{\boldsymbol{x_i}} + h\left(y\right)\frac{x_{i+1}-y}{x_{i+1}-x_i}$, and $\widehat{\boldsymbol{v}}_{\boldsymbol{x_{i+1}}} := \widehat{\boldsymbol{v}}_{\boldsymbol{x_{i+1}}} + h\left(y\right)\frac{y-x_i}{x_{i+1}-x_i}$. Observe that such interpolation preserves both the probability mass residing on y and the quantity $h\left(y\right)$. Then it is proven that for each y,

$$h\left(y\right)\left|\left(\frac{x_{i+1}-y}{x_{i+1}-x_i}\operatorname{Pois}\left(mx_i, j\right) + \frac{y-x_i}{x_{i+1}-x_i}\operatorname{Pois}\left(mx_{i+1}, j\right)\right) - \operatorname{Pois}\left(my, j\right)\right| \tag{2.10}$$

is bounded. Since fingerprint entries are closely concentrated around their expectations, bounding Expression 2.10 implies that the third condition of LP is satisfied. Intuitively, Expression 2.10 is bounded due to the sufficiently dense structure of the set X. In the next step of the construction of $\boldsymbol{v'}$, a normalization is realized to ensure that the second constraint of LP is satisfied, that is, the probability mass shared among the members of the set X and the probability mass in the empirical distribution derived from the fingerprint of frequently occurring elements add up to 1. Then it is shown that such normalization has a small effect on Expression 2.10. Therefore, $\widehat{\boldsymbol{v}}$ is in the feasible region with high probability. Intuitively, the associated histogram $h^{\widehat{\boldsymbol{v}}}$ is close to h with respect to relative earthmover distance, since for the low probability portion of the distribution, the two histograms are highly similar by construction, and for the high probability portion of the distribution, $h^{\widehat{\boldsymbol{v}}}\left(\frac{i}{m}\right) = \mathcal{F}_i$ is close to $h\left(\frac{i}{m}\right)$ because each element of the fingerprint, with high probability, has true probability close to its observed probability.

In the second part of the proof of Theorem 2.15, it is shown that for any two solutions $\boldsymbol{v}, \boldsymbol{w}$ the associated histograms $h^{\boldsymbol{v}}$ and $h^{\boldsymbol{w}}$ are close. We start with key definitions.

Definition 2.16. *For a given m, a $\beta-$bump earthmoving scheme is defined by a sequence of positive real numbers $\{c_i\}$, the bump centers, and a sequence of functions $\{f_i\} : (0,1] \rightarrow \mathbb{R}$ such that $\sum\limits_{i=0}^{\infty} f_i(x) = 1$ and for each x, and each function f_i may be expressed as a linear combination of Poisson functions $f_i(x) = \sum\limits_{j=0}^{\infty} a_{ij} \operatorname{Pois}(mx, j)$ such that $\sum\limits_{j=0}^{\infty} |a_{ij}| \leq \beta$. Given a generalized histogram h, the scheme works as follows: for each x such that $h(x) \neq 0$, and each integer $i > 0$, move $xh(x) \cdot f_i(x)$ probability mass from x to c_i. We denote the histogram resulting from this scheme by $(c, f)(h)$.*

Definition 2.17. *For given n, m, a bump earthmoving scheme (c, f) is $\epsilon-$good if for any generalized histogram h, the relative earthmover distance between h and $(c, f)(h)$ is at most ϵ.*

As these definitions hint, it is sufficient to construct a "good" bump earthmoving scheme such that, given two solutions $\boldsymbol{v}, \boldsymbol{w}$, when the scheme is applied to the associated histograms $h^{\boldsymbol{v}}, h^{\boldsymbol{w}}$, the resulting histograms $(c, f)(h^{\boldsymbol{v}})$ and $(c, f)(h^{\boldsymbol{w}})$ have small distance. Then it immediately follows that $(c, f)(h^{\boldsymbol{v}})$ and $(c, f)(h^{\boldsymbol{w}})$ have similar entropies, leading to the same conclusion about the entropies of $h^{\boldsymbol{v}}$ and $h^{\boldsymbol{w}}$ which ends the proof.

Lemma 2.18. ([20], Lemma 16) *For $n > m$, letting ς be such that $n = \varsigma m \log m$, there exists an $O\left(\sqrt{\varsigma} \cdot \max\{1, |\log \varsigma|\}\right) -$ good $m^{0.3}-$bump earthmoving scheme.*

We only present the general idea of the proof due to its laborious details which may be tiresome for a reader to follow. The building block of the construction of a "good" bump earthmoving scheme is to use two different classes of functions for $\{f_i\}$. For $i \geq \log m$ Poisson functions $\operatorname{Pois}(mx, i)$ are utilized. For $i < \log m$ Chebyshev polynomials $T_i(x)$ are employed where the i^{th} Chebyshev polynomial is the polynomial of degree i such that $T_i(\cos y) = \cos(i \cdot y)$. The sequence $\{c_i\}$ is assigned to be $\{\frac{i}{m}\}$ for

$i \geq \log m$, and $\left\{ \frac{2 \log m}{5m} \left(1 - \cos \left(\frac{5(i+1)\pi}{\log m} \right) \right) \right\}$ for $i < \log m$. Then by leveraging the fact that any two solutions $\boldsymbol{v}, \boldsymbol{w}$ must have close fingerprint expectations, it is shown that the resulting histograms $(c, f)(h^{\boldsymbol{v}})$ and $(c, f)(h^{\boldsymbol{w}})$ have small earthmoving distance.

2.1.2. Lower Bound

In the second part, we introduce the lower bound established in [19] on estimating entropy. Valiant and Valiant prove that the task of estimating entropy up to an additive error has sample complexity $\Omega\left(\frac{n}{\log n}\right)$ by constructing two probability distributions $\boldsymbol{p}^{+}$ and $\boldsymbol{p}^{-}$ with large relative earthmover distance yet having close fingerprint distributions. In other words, $\boldsymbol{p}^{+}$ and $\boldsymbol{p}^{-}$ are built such that the difference, $|H(\boldsymbol{p}^{+}) - H(\boldsymbol{p}^{-})|$, is big enough to distinguish $\boldsymbol{p}^{+}$ from $\boldsymbol{p}^{-}$, however, no algorithm can differentiate between the fingerprint obtained from $\boldsymbol{p}^{+}$ and the fingerprint obtained from $\boldsymbol{p}^{-}$ by drawing $o\left(\frac{n}{\log n}\right)$ samples. In addition, Valiant and Valiant provide two new central limit theorems (CLT) for establishing the lower bound. We choose to avoid the details of the proof of the central limit theorems in order to preserve the smoothness of the explanation.

Definition 2.19. *The generalized multinomial distribution parameterized by a nonnegative matrix ρ each of whose rows sum to at most 1, is denoted M^{ρ}, and is defined by the following random process: for each row $\rho(i, \cdot)$ of matrix ρ, interpret it as a probability distribution over the columns of ρ — including, if $\sum_{j=1}^{m} \rho(i, j) < 1$, an "invisible" column 0 — and draw a column index from this distribution; return a row vector recording the total number of samples falling into each column (the histogram of the samples).*

The generalized multinomial distribution is employed to capture the fingerprint distribution of a probability distribution. We introduce the first central limit theorem that relates the sum of independent distributions to a Gaussian distribution with respect to earthmover distance.

Theorem 2.20. ([19], Theorem 2) *Given n independent distributions $\{Z_i\}$ of mean 0 in $\mathbb{R}^m$ and a bound β such that $\|Z_i\| < \beta$ for any i and any sample, the earthmover* *distance between $\sum_{i=1}^{n} Z_i$ and the normal distribution of corresponding mean* (0) *and covariance is at most* $\beta m(2.7 + 0.83 \log n)$.

The second central limit theorem approximates a generalized multinomial distribution by a Gaussian distribution with respect to statistical distance. Note that the statistical distance between a generalized multinomial distribution and a Gaussian distribution is 1 since the first is discrete but the second is continuous. Therefore, Gaussian distribution is discretized by rounding to the nearest lattice points.

Definition 2.21. *The $m-$dimensional discretized Gaussian distribution, with mean μ and covariance matrix Σ, denoted $\mathcal{N}^{disc}(\mu, \Sigma)$, is the distribution with support $\mathbb{Z}^m$ obtained by picking a sample according to the Gaussian $\mathcal{N}(\mu, \Sigma)$, then rounding each coordinate to the nearest integer.*

Theorem 2.22. ([19], Theorem 4) *Given a generalized multinomial distribution M^ρ, with m dimensions and n rows, let μ denote its mean and Σ denote its covariance matrix, then*

$$\mathrm{d}_{\mathrm{TV}}\left(M_\rho, \mathcal{N}^{disc}(\mu, \Sigma)\right) \leq \frac{m^{4/3}}{\sigma^{1/3}} \cdot 2.2 \cdot (3.1 + 0.83 \log n)^{2/3},$$

where σ^2 is the minimum eigenvalue of Σ.

We present how distributions $\boldsymbol{p}^+$ and $\boldsymbol{p}^-$ are constructed.

Definition 2.23. *Given a real number $\phi \in \left(0, \frac{1}{4}\right)$, consider the degree $\log m + 2$ polynomial $M_{\log m, \phi}(x) := -\left(x - \phi\frac{1}{\log m}\right)\left(x - 2\phi\frac{1}{\log m}\right) L_{\log m}(x)$ such that $L_j(x) = \frac{e^x}{j!}\frac{d^j}{dx^j}\left(e^{-x}x^j\right)$ is the j^{th} Laguerre polynomial. Let $v(x)$ be the function that takes value $1/M'_{\log m, \phi}(x)$ for every x where $M_{\log m, \phi}(x) = 0$, and is 0 otherwise, where M' is the derivative of M. Define the distributions $\boldsymbol{p}^+_{\boldsymbol{\log m, \phi}}, \boldsymbol{p}^-_{\boldsymbol{\log m, \phi}}$ such that for each x where $v(x) > 0$, the distribution $\boldsymbol{p}^+_{\boldsymbol{\log m, \phi}}$ contains $v(x)e^{x/32}$ probability mass at probability $\frac{1}{32m}x$, and for each x where $v(x) < 0$ the distribution $\boldsymbol{p}^-_{\boldsymbol{\log m, \phi}}$ contains $|v(x)|e^{x/32}$ probability mass at probability $\frac{1}{32m}x$, where each distribution is then normalized to have total probability mass 1.*

Observe that since the probability of each element of either $\boldsymbol{p}^{+}_{\log m,\phi}$ or $\boldsymbol{p}^{-}_{\log m,\phi}$ is defined to be at least $\frac{\phi}{32m\log m}$, both distributions have support at most $\frac{32}{\phi}m\log m$. Thus, the connection between the sample and domain sizes necessary for the lower bound is formed. The second condition on $\boldsymbol{p}^{+}_{\log m,\phi}$, $\boldsymbol{p}^{-}_{\log m,\phi}$ is that the difference $|H\left(\boldsymbol{p}^{+}_{\log m,\phi}\right) - H\left(\boldsymbol{p}^{-}_{\log m,\phi}\right)|$ is sufficiently big. This can be achieved by proving that $\boldsymbol{p}^{+}_{\log m,\phi}$, $\boldsymbol{p}^{-}_{\log m,\phi}$ are close in the relative earthmover distance to two different distributions $\boldsymbol{q}^{+}, \boldsymbol{q}^{-}$, respectively, such that $H\left(\boldsymbol{q}^{+}\right)$ is distant from $H\left(\boldsymbol{q}^{-}\right)$.

Lemma 2.24. ([19], Lemma 13) *Distributions* $\boldsymbol{p}^{+}_{\log m,\phi}$, $\boldsymbol{p}^{-}_{\log m,\phi}$ *are* $O\left(\phi|\log\phi|\right)$ −*close, respectively, in the relative earthmover distance to the uniform distributions on* $\frac{32}{\phi}m\log m$ *and* $\frac{16}{\phi}m\log m$ *elements.*

Note that the relative earthmover distance, therefore by Fact 2.10, the difference of the entropies is $H\left(\mathcal{U}\left(\left[\frac{32}{\phi}m\log m\right]\right)\right) - H\left(\mathcal{U}\left(\left[\frac{16}{\phi}m\log m\right]\right)\right) = 1$, where $\mathcal{U}\left([n]\right)$ denotes the uniform distribution with support size equal to n. That is, $\boldsymbol{p}^{+}_{\log m,\phi}$ and $\boldsymbol{p}^{-}_{\log m,\phi}$ are distinguishable. The third condition is that $\boldsymbol{p}^{+}_{\log m,\phi}$, $\boldsymbol{p}^{-}_{\log m,\phi}$ have fingerprint distributions $\mathcal{F}_{\boldsymbol{p}^{+}}$, $\mathcal{F}_{\boldsymbol{p}^{-}}$, respectively, such that the statistical distance between $\mathcal{F}_{\boldsymbol{p}^{+}}$ and $\mathcal{F}_{\boldsymbol{p}^{-}}$ is small. Similarly, it is shown that $\mathcal{F}_{\boldsymbol{p}^{+}}$, $\mathcal{F}_{\boldsymbol{p}^{-}}$ are approximated by two statistically close, discretized Gaussian distributions $\mathcal{N}^{disc}_{+}$, $\mathcal{N}^{disc}_{-}$, respectively. First, we state a weaker result.

Lemma 2.25. ([19], Lemma 16) *For any* i*, the* i^{th} *fingerprint expectations for distributions* $\boldsymbol{p}^{+}_{\log m,\phi}, \boldsymbol{p}^{-}_{\log m,\phi}$ *are equal to within* $o(1)$.

It is necessary for $\boldsymbol{p}^{+}_{\log m,\phi}$, $\boldsymbol{p}^{-}_{\log m,\phi}$ to have bigger variance in both directions due to obtaining better bounds when aforementioned central limit theorems are applied. Therefore, $\boldsymbol{p}^{+}_{\log m,\phi}$, $\boldsymbol{p}^{-}_{\log m,\phi}$ are modified to get "fat" distributions $\boldsymbol{p}^{\boldsymbol{F}_{+}}_{\log m,\phi}$, $\boldsymbol{p}^{\boldsymbol{F}_{-}}_{\log m,\phi}$ such that both are statistically close to their "thin" counterparts, respectively. A "fat" distribution is constructed as follows:

Definition 2.26. *Define the fattening operator* F *that, given a histogram* h*, constructs a new histogram* h^{F} *as follows:*

- *Provisionally set* $h^F(x) = \left(1 - \frac{\log m - 1}{2\log^2 m}\right) h(x)$ *for each* x*;*
- *For each integer* $i \in \{1, \ldots, \log m\}$*, increment* $h^F\left(\frac{i}{m}\right) \leftarrow h^F\left(\frac{i}{m}\right) + \frac{m}{\log^3 m}$.

Note that operator F returns a proper probability distribution and preserves the previous upper bound on support size since no element with probability less than $1/m$ is added to the support. Intuitively, it also "fattens" a distribution because the number of low probability elements is increased substantially. Moreover, operator F does not negatively affect the bounds of Lemma 2.25 since both distributions $\boldsymbol{p}^{+}_{\boldsymbol{\log m,\phi}}$, $\boldsymbol{p}^{-}_{\boldsymbol{\log m,\phi}}$ are modified identically. We are ready to state the main result.

Proposition 2.27. ([19], Proposition 21) *For a positive constant* $\phi < 1/4$*, the statistical distance between the distribution of* $\text{Pois}(m)$*−sample fingerprints from* $\boldsymbol{p}^{\boldsymbol{F}_+}_{\boldsymbol{\log m,\phi}}$ *and* $\boldsymbol{p}^{\boldsymbol{F}_-}_{\boldsymbol{\log m,\phi}}$ *goes to* 0 *as* m *goes to infinity.*

Recall that Theorem 2.22 is utilized to approximate the fingerprint distributions of $\boldsymbol{p}^{\boldsymbol{F}_+}_{\boldsymbol{\log m,\phi}}$, $\boldsymbol{p}^{\boldsymbol{F}_-}_{\boldsymbol{\log m,\phi}}$ by two statistically close, discretized Gaussian distributions, respectively. We only explain the necessary conditions for Theorem 2.22, skipping the details of its application.

The condition to be satisfied is that fingerprint distributions $\mathcal{F}_{\boldsymbol{p}^{\boldsymbol{F}_+}}$, $\mathcal{F}_{\boldsymbol{p}^{\boldsymbol{F}_-}}$ obtained from distributions $\boldsymbol{p}^{\boldsymbol{F}_+}_{\boldsymbol{\log m,\phi}}$, $\boldsymbol{p}^{\boldsymbol{F}_-}_{\boldsymbol{\log m,\phi}}$, respectively, have close variance and covariance. This is proven by using the result in Lemma 2.25, that is, fingerprint distributions having close expectations. Recall that for a histogram h, expectation of the i^{th} fingerprint entry is $\mathbf{E}[\mathcal{F}_i] = \sum_{x:h(x)\neq 0} h(x) \cdot \text{Pois}(mx, i)$, and covariance of two random variables X, Y is defined as $\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$. Combining with Equation 2.9, it follows that covariance of the i^{th} and j^{th} fingerprint entries, for $i \neq j$, equals $\mathbf{Cov}[\mathcal{F}_i, \mathcal{F}_j] = \sum_{x:h(x)\neq 0} -h(x) \cdot \text{Pois}(xm, i)\,\text{Pois}(xm, j)$. After the simplification,

$$\text{Pois}(xm, i)\,\text{Pois}(xm, j) = \frac{(xm)^{i+j}\, e^{-2xm}}{i!j!} = 2^{-(i+j)} \binom{i+j}{i} \text{Pois}(2xm, i+j)\,.$$

Recall that variance of a random variable X is $\mathbf{Var}\left[X\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}^2\left[X\right]$. Then for the i^{th} fingerprint entry, $\mathbf{Var}\left[\mathcal{F}_i\right] = \sum_{x:h(x)\neq 0} h\left(x\right) \cdot \left(\mathrm{Pois}\left(mx, i\right) - \mathrm{Pois}\left(mx, i\right)^2\right)$. Note that $\mathrm{Pois}\left(mx, i\right)^2 = 2^{-2i}\binom{2i}{i} \cdot \mathrm{Pois}\left(2mx, 2i\right)$. The following relates Poisson functions with different parameters.

Lemma 2.28. ([19], Lemma 20) *For any $\epsilon > 0$ and integer $i \geq 0$, one may approximate* $\mathrm{Pois}\left(2x, i\right)$ *as a linear combination* $\sum_{j=0}^{\infty} \alpha\left(j\right) \mathrm{Pois}\left(x, j\right)$ *such that*

1. *For all $x \geq 0$, $|\,\mathrm{Pois}\left(2x, i\right) - \sum_{j=0}^{\infty} \alpha\left(j\right) \mathrm{Pois}\left(x, j\right)| \leq \epsilon$; and*
2. *$\sum_{j=0}^{\infty} |\alpha\left(j\right)| \leq \frac{1}{\epsilon} \cdot 200 \max\left\{\sqrt[4]{i}, 24 \log^{3/2} \frac{1}{\epsilon}\right\}$.*

Therefore, both variance and covariance of fingerprint distributions of $\boldsymbol{p}_{\mathbf{log}\,\boldsymbol{m},\boldsymbol{\phi}}^{\boldsymbol{F}_+}$ and $\boldsymbol{p}_{\mathbf{log}\,\boldsymbol{m},\boldsymbol{\phi}}^{\boldsymbol{F}_-}$ can be expressed as linear combinations of Poisson functions. It is already known that fingerprint distributions of $\boldsymbol{p}_{\mathbf{log}\,\boldsymbol{m},\boldsymbol{\phi}}^{\boldsymbol{F}_+}$, $\boldsymbol{p}_{\mathbf{log}\,\boldsymbol{m},\boldsymbol{\phi}}^{\boldsymbol{F}_-}$ have close expectations which themselves are expressed as linear combinations of Poisson functions. Then, at least intuitively, fingerprint distributions of $\boldsymbol{p}_{\mathbf{log}\,\boldsymbol{m},\boldsymbol{\phi}}^{\boldsymbol{F}_+}$, $\boldsymbol{p}_{\mathbf{log}\,\boldsymbol{m},\boldsymbol{\phi}}^{\boldsymbol{F}_-}$ have similar variance and covariance.

Theorem 2.29. ([19], Theorem 1) *For any positive constant $\phi < 1/4$ there exists a pair of distributions $\mathbf{p}^+, \boldsymbol{p}^-$ that are $O\left(\phi|\log\phi|\right) -$ close in the relative earthmover distance, respectively, to the uniform distributions on n and $n/2$ elements, but which are indistinguishable to $m = \frac{\phi}{32} \cdot \frac{n}{\log n} -$sample testers.*

2.1.3. Upper Bound II

In the final part of this section, we present an improved upper bound on estimating entropy up to an additive Δ. The bounds given in [19, 20] are matching in their dependence on n, whereas for the dependence on Δ this is not the case. The estimator constructed in [20] has sample complexity $O\left(\frac{n}{\Delta^2 \log n}\right)$, while the lower bound established in [19] is $\Omega\left(\frac{n}{\Delta \log n}\right)$, which leaves open the question of error decrease rate. The problem is resolved in [21] by constructing an optimal estimator which estimates the entropy of a distribution to within additive accuracy Δ, with probability at least $1 - o\left(poly\left(n\right)\right)$, given $O\left(\frac{n}{\Delta \log n}\right)$ independent samples from a distribution with support size at most n. Differently than the previous one which is based on a canonical approach, Valiant and Valiant construct an estimator focusing directly on entropy.

Definition 2.30. *A symmetric property π is **linear** if there exists some function $f_\pi : [0,1] \to \mathbb{R}$, denoted as **characteristic function** of π, such that for any distribution $\boldsymbol{p}$ with histogram $h_{\boldsymbol{p}}$, $\pi(\boldsymbol{p}) = \sum\limits_{x:h_{\boldsymbol{p}}(x)\neq 0} h_{\boldsymbol{p}}(x) f_\pi(x)$.*

Observe that Shannon entropy is a linear property, and its characteristic function is $f(x) = x|\log x|$. The new estimator is based on approximating the characteristic function of entropy as a linear combination of Poisson functions. The following clarifies the motivation behind this approach. Assume there exists a sequence of coefficients $\{\beta_i\}$ such that for all $x \in (0,1]$, $\sum_{i=1}^{\infty} \beta_i \operatorname{Pois}(mx, i) = x|\log x| = f(x)$. Then,

$$\begin{aligned}\sum_{x:h(x)\neq 0} h(x)f(x) &= \sum_{x:h(x)\neq 0} h(x) \sum_{i\geq 1} \beta_i \operatorname{Pois}(mx,i) \\ &= \sum_{i\geq 1} \beta_i \sum_{x:h(x)\neq 0} h(x) \operatorname{Pois}(xm,i) \\ &= \sum_{i\geq 1} \beta_i \, \mathbf{E}[\mathcal{F}_i] = \mathbf{E}\left[\sum_{i\geq 1} \beta_i \mathcal{F}_i\right]. \end{aligned} \tag{2.11}$$

That is, the quantity $\sum\limits_{i\geq 1} \beta_i \mathcal{F}_i$ is an unbiased estimator for entropy. Recall that for each i, a fingerprint entry $\mathcal{F}_i$ is concentrated around its expectation. Then, roughly, for $\sum\limits_{i\geq 1} \beta_i \mathcal{F}_i$ having relatively small variance one needs the coefficients $\{\beta_i\}$ to be small comparing to $1/\sqrt{m}$. However, instead of approximating the characteristic function $f(x) = x|\log x|$ directly, the function $\frac{f(x)}{x} = |\log x|$ is expressed as a linear combination of Poisson functions, $\sum_{i=0}^{\infty} z_i \operatorname{Pois}(mx, i)$. Observe that these approaches are equivalent in the sense that $\beta_i = \frac{i}{m} \cdot z_{i-1}$, since $x \operatorname{Pois}(mx,i) = \operatorname{Pois}(mx, i+1)\frac{i+1}{m}$. The following formalizes the relationship between the magnitudes of coefficients, error in approximating $|\log x|$ and the estimator defined above.

Proposition 2.31. ([21], Proposition 17) *Given integers m, n, and a set of coefficients $z_0, z_1, \ldots$ such that if for positive real numbers a, b, c the following conditions hold:*

1. $\left| |\log x| - \sum\limits_{i=0}^{\infty} z_i \operatorname{Pois}(mx, i) \right| < a + \frac{b}{x}$,

2. *for all* $j \geq 1$ *let* $\beta_j = \frac{j}{m} z_{j-1}$ *with* $\beta_0 = 0$*, then for any* j,l *such that* $|j - l| \leq \sqrt{j} \log m$ *we have* $|\beta_j - \beta_l| \leq c\sqrt{\frac{j}{m}}$.

Then the estimator described in Equation 2.11 estimates entropy with error at most $a + bn + c \log m$*, with probability at least* $1 - o\left(1/poly\left(m\right)\right)$ *when given a fingerprint derived from a set of* m *independent samples chosen from a distribution with support size at most* n*.*

The task of finding such coefficients $\{z_i\}$ is realized via linear programming. A linear program is constructed with constraints describing the conditions of Proposition 2.31 and with the objective function minimizing error in the estimation.

```
Algorithm: ESTIMATOR II
  • Fix $m = O\left(\frac{n}{\Delta \log n}\right)$. (For details see [21].)
  • Draw $M \sim \mathrm{Pois}\left(m\right)$ independent samples $X_1, \ldots, X_M$ from $\boldsymbol{p}$.
  • Construct the linear program as in Definition 18 [21], corresponding to $\mathcal{F}$.
  • Find a solution $\{z_i\}$ to the the linear program.
  • Calculate the coefficients $\{\beta_i\}$.
  • Output $\sum_{i \geq 1} \beta_i \mathcal{F}_i$.
```

Figure 2.2. Linear Estimator of Shannon Entropy.

Aside from employing linear programming to find convenient coefficients $\{z_i\}$, Valiant and Valiant explicitly construct an optimal estimator such that given $O\left(\frac{n}{\Delta \log n}\right)$ independent samples from a distribution with support size at most n, it estimates entropy of a distribution to within additive accuracy Δ, with probability at least $1 - o\left(poly\left(n\right)\right)$. We briefly describe how an optimal estimator is constructed. Recall that the objective is to approximate the function $\log x$ as a linear combination of Poisson functions, $\sum_{i=0}^{\infty} z_i \, \mathrm{Pois}(mx, i)$. A straightforward choice for coefficients $\{z_i\}$ is the sequence $\left\{\log \frac{i}{m}\right\}$. Note that $\log \frac{i}{m}$ is a "plug-in" estimator for entropy. The following lemma bounds the precision of any "plug-in" estimator.

Lemma 2.32. ([21], Lemma 19) *Given a function $f : \mathbb{R} \to \mathbb{R}$ whose fourth derivative at x is bounded in magnitude by $\frac{\alpha}{x^4}$ for $x \geq 1$ and by α for $x \leq 1$, and whose third derivative at x is bounded by $\frac{\alpha}{x^3}$, then for any real x, $\sum_{i=0}^{\infty} f(i)\, \mathrm{Pois}(x,i)$ is within $O\left(\frac{\alpha}{x^2}\right)$ of $f(x) + \frac{1}{2} x f''(x)$.*

For a "plug-in" estimator $\log \frac{i}{m}$, this lemma suggests that

$$\log x - \sum_{i=0}^{\infty} \log(i/m)\, \mathrm{Pois}(mx, i) = \frac{-1}{2 \ln 2 \cdot mx} + O\left(\frac{1}{m^2 x^2}\right). \tag{2.12}$$

Observe that for a high probability, error of approximation is small, whereas for a low probability such as $x \leq \frac{1}{m}$, the right-hand side of Equation 2.12 becomes unbounded. In other words, "plug-in" estimator is satisfactory for high probability portion of a distribution, however, it behaves poorly for low probability portion of a distribution. Similar techniques are used as in bump earthmoving scheme described in Subsection 2.1.1 to resolve the issue. We only give an outline of the proof due to its laborious details.

Two different functions are used for approximating $\log x$ as a linear combination of Poisson functions. For probabilities $x \geq O(\log m)$ "plug-in" estimator $\log \frac{i}{m}$ and for probabilities $x < O(\log m)$ a function of Chebyshev polynomials referred as the Chebyshev bumps are utilized. The next step is to establish a Chebyshev bump version of Lemma 2.32 and to show that the Chebyshev bumps can be expressed as a linear combination of Poisson functions with relatively small coefficients. The proof ends by applying Proposition 2.31.

2.2. Rényi Entropy

We investigate Rényi entropy, first introduced by Alfréd Rényi [25] which is a popular generalization of Shannon entropy. It is defined as follows:

Definition 2.33. *Let $\alpha \geq 0$ be a real number. The Rényi entropy of order α of a distribution $\boldsymbol{p}$, denoted by $H_\alpha(\boldsymbol{p})$, is* *for* $\alpha \neq 1$

$$H_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \left(\sum_x \boldsymbol{p}[x]^\alpha \right), \tag{2.13}$$

and for $\alpha = 1$

$$H_1(\boldsymbol{p}) = \lim_{\alpha \to 1} H_\alpha(\boldsymbol{p}) \tag{2.14}$$

For $\alpha = 0$, $H_0(\boldsymbol{p}) = \log \operatorname{supp}(\boldsymbol{p})$, where $\operatorname{supp}(\boldsymbol{p})$ denotes the support size of the distribution $\boldsymbol{p}$. For $\alpha = 1$, $H_1(\boldsymbol{p}) = H(\boldsymbol{p})$, that is, Rényi entropy becomes Shannon entropy, as easily derived via L'Hôpital's rule. For $\alpha = \infty$, $H_\alpha(\boldsymbol{p})$ is the min-entropy $H_\infty(\boldsymbol{p})$, where by definition $H_\infty(\boldsymbol{p}) = -\log \max_i \boldsymbol{p}[i]$.

Rényi entropy has many applications. Particularly, $H_2(\boldsymbol{p})$ is used for measuring the quality of random number generators [26], for testing the closeness of probability distributions [27, 28], for characterizing the number of reads needed to reconstruct a DNA sequence [29], etc.

Recall that $\Theta\left(\frac{n}{\log n}\right)$ samples are necessary and sufficient for estimating Shannon entropy, which are only better by a polylogarithmic factor than $\Theta\left(n \log^2 n\right)$, a trivial upper bound for this task. Thus, being a generalization of Shannon entropy, determining the complexity of estimating Rényi entropy becomes additionally intriguing. Acharya, Orlitsky, Suresh, and Tyagi [7] provide near-optimal upper and lower bounds for three different cases of α; it is shown that to estimate Rényi entropy to within an additive error one requires (i) for $\alpha < 1$, super-linear, roughly $n^{1/\alpha}$ samples (ii) for noninteger $\alpha > 1$, near-linear, roughly n samples (iii) for integer $\alpha > 1$, sub-linear, $\Theta\left(n^{1-1/\alpha}\right)$ samples. Note that in the case of $\alpha > 1$ being integer, estimating Rényi entropy becomes substantially easier than estimating Shannon entropy.

2.2.1. Upper Bounds

In the first part we illustrate the upper bounds established in [7]. Defining the α^{th} moment of $\boldsymbol{p}$ as $\mathcal{M}_\alpha(\boldsymbol{p}) = \sum_{i=1}^{n} (\boldsymbol{p}[i])^\alpha$, Rényi entropy can be expressed as $H_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \mathcal{M}_\alpha(\boldsymbol{p})$. Observe that estimating $H_\alpha(\boldsymbol{p})$ to an additive accuracy of $\pm\Delta$ is equivalent to estimating $\mathcal{M}_\alpha(\boldsymbol{p})$ to a multiplicative accuracy of $2^{\pm\Delta(1-\alpha)}$. Acharya *et al.* construct two different multiplicative estimators, denoted by $\widehat{\mathcal{M}^e_\alpha}$ and $\widehat{\mathcal{M}^u_\alpha}$, the first for $\alpha \notin \mathbb{Z}$ and the latter for $\alpha \in \mathbb{Z}$.

To simplify the analysis, Poisson sampling technique is utilized as in the case of estimating Shannon entropy whose explicit description is given in Section 2.1. Recall that in the Pois(m)-sampling scheme N_i/m is an unbiased estimator for $\boldsymbol{p}[i]$ where N_i denotes the multiplicity of an element i. We define the aforementioned estimators $\widehat{\mathcal{M}^e_\alpha}$, $\widehat{\mathcal{M}^u_\alpha}$.

Definition 2.34. *For $\alpha \notin \mathbb{Z}$, let the empirical estimator for $\mathcal{M}_\alpha(\boldsymbol{p})$, denoted by $\widehat{\mathcal{M}^e_\alpha}$, be*

$$\widehat{\mathcal{M}^e_\alpha} = \sum_{i\in[n]} \left(\frac{N_i}{m}\right)^\alpha. \tag{2.15}$$

Observe that $\widehat{\mathcal{M}^e_\alpha}$ is biased. For $\alpha \in \mathbb{Z}^+$, let $n^{\underline{\alpha}} = n\cdot(n-1)\cdots(n-\alpha+1)$ denote the α^{th} *falling power* of n.

Definition 2.35. *For integer $\alpha > 1$, let the bias-corrected estimator for $\mathcal{M}_\alpha(\boldsymbol{p})$, denoted by $\widehat{\mathcal{M}^u_\alpha}$, be*

$$\widehat{\mathcal{M}^u_\alpha} = \sum_{i\in[n]} \frac{N_i^{\underline{\alpha}}}{m^\alpha}. \tag{2.16}$$

We demonstrate the central lemma for establishing upper bounds, which essentially constructs an error reduction algorithm to increase the accuracy of an estimator given certain bounds on its bias and variance.

Lemma 2.36. ([7], Lemma 6) *For $M \sim \text{Pois}(m)$, let the estimator $\widehat{\mathcal{M}_\alpha}$ have bias and variance satisfying*

$$\begin{aligned}\left|\mathbf{E}\left[\widehat{\mathcal{M}_\alpha}\right] - \mathcal{M}_\alpha(\boldsymbol{p})\right| &\leq \frac{\gamma}{2}\mathcal{M}_\alpha(\boldsymbol{p}),\\ \mathbf{Var}\left[\widehat{\mathcal{M}_\alpha}\right] &\leq \frac{\gamma^2}{12}\mathcal{M}_\alpha(\boldsymbol{p})^2.\end{aligned}$$

Then, there exists an estimator $\widehat{\mathcal{M}'}_\alpha$ that uses $O\left(m \log\left(1/\delta\right)\right)$ samples, and ensures

$$\Pr\left[\left|\widehat{\mathcal{M}'}_\alpha - \mathcal{M}_\alpha(\boldsymbol{p})\right| > \gamma\mathcal{M}_\alpha(\boldsymbol{p})\right] \leq \delta.$$

Proof. By Chebyshev's inequality,

$$\begin{aligned}&\Pr\left[\left|\widehat{\mathcal{M}_\alpha} - \mathcal{M}_\alpha(\boldsymbol{p})\right| > \gamma\mathcal{M}_\alpha(\boldsymbol{p})\right]\\ &\leq \Pr\left[\left|\widehat{\mathcal{M}_\alpha} - \mathbf{E}\left[\widehat{\mathcal{M}_\alpha}\right]\right| + \left|\mathbf{E}\left[\widehat{\mathcal{M}_\alpha}\right] - \mathcal{M}_\alpha(\boldsymbol{p})\right| > \gamma\mathcal{M}_\alpha(\boldsymbol{p})\right]\\ &\leq \Pr\left[\left|\widehat{\mathcal{M}_\alpha} - \mathbf{E}\left[\widehat{\mathcal{M}_\alpha}\right]\right| > \frac{\gamma}{2}\mathcal{M}_\alpha(\boldsymbol{p})\right]\\ &\leq \frac{4\,\mathbf{Var}\left[\widehat{\mathcal{M}_\alpha}\right]}{\gamma^2\widehat{\mathcal{M}_\alpha}^2} \leq \frac{1}{3}.\end{aligned}$$

$\mathcal{M}_\alpha(\boldsymbol{p})$ is estimated in t independent rounds, and $\widehat{M}'_\alpha$ is assigned to be the sample median of these rounds. More specifically, let $\widehat{\mathcal{M}_i}$ denote the result of round i, and let $\mathbb{1}_{E_i}$ be the indicator function of the event $E_i = \left\{\left|\widehat{\mathcal{M}_i} - \mathcal{M}_\alpha(\boldsymbol{p})\right| > \gamma\mathcal{M}_\alpha(\boldsymbol{p})\right\}$. Then the expectation satisfies $\mathbf{E}\left[\mathbb{1}_{E_i}\right] \leq 1/3$, and by the Hoeffding bound,

$$\Pr\left[\sum_{i=1}^{t}\mathbb{1}_{E_i} > \frac{t}{2}\right] \leq e^{-t/18}.$$

To reduce the probability of error to δ, set $t = 18\log\left(1/\delta\right)$. What this means is that with probability at least $1-\delta$, the majority of the rounds, therefore, the sample median $\widehat{\mathcal{M}'}_\alpha$ satisfies the condition $\left|\widehat{\mathcal{M}_i} - \mathcal{M}_\alpha(\boldsymbol{p})\right| \leq \gamma\mathcal{M}_\alpha(\boldsymbol{p})$. □

The next step is to bound the bias and variance of both estimators, $\widehat{\mathcal{M}^e_\alpha}$, $\widehat{\mathcal{M}^u_\alpha}$, in order to apply Lemma 2.36, which finalizes the proof. Note that independence gained due to Poisson sampling simplifies the analysis in bounding variance. We only state related results regarding the bias-corrected estimator.

Lemma 2.37. ([7], Lemma 2) *Let* $X \sim \mathrm{Pois}(\lambda)$. *Then, for all* $r \in \mathbb{N}$

$$\begin{aligned}\mathbf{E}[X^{\underline{r}}] &= \lambda^r,\\ \mathbf{Var}[X^{\underline{r}}] &\leq \lambda^r \left((\lambda + r)^r - \lambda^r\right).\end{aligned}$$

Proof. The expected value is

$$\mathbf{E}[X^{\underline{r}}] = \sum_{i=0}^{\infty} \mathrm{Pois}(\lambda, i) \cdot i^{\underline{r}} = \sum_{i=r}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \cdot \frac{i!}{(i-r)!} = \lambda^r \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} = \lambda^r.$$

We bound the following

$$\begin{aligned}\mathbf{E}[(X^{\underline{r}})^2] &= \sum_{i=0}^{\infty} \mathrm{Pois}(\lambda, i) \cdot (i^{\underline{r}})^2\\ &= \sum_{i=r}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \cdot \frac{(i!)^2}{(i-r)!^2}\\ &= \lambda^r \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \, (i+r)^{\underline{r}}\\ &= \lambda^r \, \mathbf{E}[(X+r)^{\underline{r}}]\\ &= \lambda^r \, \mathbf{E}\left[\prod_{j=1}^{r} [(X + 1 - j) + r]\right]\\ &\leq \lambda^r \, \mathbf{E}\left[\sum_{j=0}^{r} \binom{r}{j} X^{\underline{j}} \cdot r^{r-j}\right]\\ &= \lambda^r \sum_{j=0}^{r} \binom{r}{j} \lambda^j \cdot r^{r-j}\\ &= \lambda^r \, (\lambda + r)^r .\end{aligned}$$

Hence,

$$\mathbf{Var}[X^{\underline{r}}] = \mathbf{E}[(X^{\underline{r}})^2] - \mathbf{E}[X^{\underline{r}}]^2 \leq \lambda^r \left((\lambda + r)^r - \lambda^r\right). \qquad \square$$

Finally, we state the upper bound results for estimating Rényi entropy.

Theorem 2.38. ([7], Theorem 9) *For an integer $\alpha > 1$, any $\Delta > 0$, and $0 < \delta < 1$, there exists an algorithm estimating with probability at least $1 - \delta$ the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with $O\left(\dfrac{n^{1-1/\alpha}}{(1 - 2^{(1-\alpha)\Delta})^2} \log \dfrac{1}{\delta}\right)$ samples.*

```
Algorithm: ESTIMATOR III
  • Fix $\gamma = 1 - 2^{(1-\alpha)\Delta}$ and $m = O\left(\frac{n^{1-1/\alpha}}{\gamma^2}\right)$. (For details see [7].)
  • Repeat the following for $t = \lceil 18 \log \frac{1}{\delta} \rceil$ independent rounds.
      · Draw $M \sim \mathrm{Pois}\,(m)$ independent samples $X_1, \ldots, X_M$ from $\boldsymbol{p}$.
      · Compute the multiplicity $N_i$ based on the samples $X_1, \ldots, X_M$ for $1 \leq i \leq n$.
      · Set $\widehat{\mathcal{M}}_j = \sum_{i \in [n]} \frac{N_i^\alpha}{m^\alpha}$ for round $1 \leq j \leq t$.
  • Output $\frac{1}{1-\alpha} \log \widehat{\mathcal{M}}_\alpha$ where $\widehat{\mathcal{M}}_\alpha$ is the median of the sequence $\{\widehat{\mathcal{M}}_j\}$.
```

Figure 2.3. SAMP Estimator of Rényi Entropy (of integer degree $\alpha > 1$)

Theorem 2.39. ([7], Theorem 7) *For $\alpha > 1$, $\Delta > 0$, and $0 < \delta < 1$, there exists an algorithm estimating with probability at least $1 - \delta$ the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with $O\left(\dfrac{n}{\gamma^{\max\{4, 1/(\alpha-1)\}}} \log \dfrac{1}{\delta}\right)$ samples where $\gamma = 1 - 2^{(1-\alpha)\Delta}$.*

Theorem 2.40. ([7], Theorem 8) *For $\alpha < 1$, $\Delta > 0$, and $0 < \delta < 1$, there exists an algorithm estimating with probability at least $1 - \delta$ the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with $O\left(\dfrac{n^{1/\alpha}}{\gamma^{\max\{4, 2/\alpha\}}} \log \dfrac{1}{\delta}\right)$ samples where $\gamma = 1 - 2^{(\alpha-1)\Delta}$.*

```
Algorithm: ESTIMATOR IV
• Fix $\gamma = 1 - 2^{(1-\alpha)\Delta}$ and $m = O\left(\frac{n}{\gamma^{\max\{4,1/(\alpha-1)\}}}\right)$. (For details see [7].)
• Repeat the following for $t = \lceil 18 \log \frac{1}{\delta} \rceil$ independent rounds.
    · Draw $M \sim \mathrm{Pois}\,(m)$ independent samples $X_1, \ldots, X_M$ from $\boldsymbol{p}$.
    · Compute the multiplicity $N_i$ based on the samples $X_1, \ldots, X_M$ for $1 \leq i \leq n$.
    · Set $\widehat{\mathcal{M}_j} = \sum_{i \in [n]} \left(\frac{N_i}{m}\right)^{\alpha}$ for round $1 \leq j \leq t$.
• Output $\frac{1}{1-\alpha} \log \widehat{\mathcal{M}_\alpha}$ where $\widehat{\mathcal{M}_\alpha}$ is the median of the sequence $\{\widehat{\mathcal{M}_j}\}$.
```

Figure 2.4. SAMP Estimator of Rényi Entropy (of noninteger degree $\alpha > 1$)

We skip the algorithmic view of additively estimating the Rényi entropy for $\alpha < 1$, since it is easily obtained by setting $m = O\left(\frac{n^{1/\alpha}}{\left(1-2^{(\alpha-1)\Delta}\right)^{\max\{4,2/\alpha\}}}\right)$ in Figure 2.4.

Very recently, Acharya, Orlitsky, Suresh, and Tyagi [30] have improved the upper bound results for additively estimating the Rényi entropy for noninteger values of α. For $\alpha < 1$ and noninteger $\alpha > 1$, they construct algorithms additively estimating (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ using $O\left(\frac{n^{1/\alpha}}{\log n}\right)$ and $O\left(\frac{n}{\log n}\right)$ samples, respectively. Both algorithms employ empirical and bias-corrected estimators, $\widehat{\mathcal{M}^e_\alpha}$ and $\widehat{\mathcal{M}^u_\alpha}$, and achieve a factor of $\log n$ improvement due to exploiting the polynomial approximation technique. In particular, for domain elements i with $N_i > \tau$, the empirical estimator $\widehat{\mathcal{M}^e_\alpha}$ is utilized, where $\tau = O\left(\log n\right)$. On the other hand, for domain elements i with $N_i \leq \tau$, the task of estimating $H_\alpha(\boldsymbol{p})$ is conducted in two steps. Firstly, the function x^α is approximated by an integer $d-$degree polynomial $p\left(x\right) = \sum_{j=0}^{d} c_j x^j$. Secondly, the quantity $\left(\boldsymbol{p}\left[i\right]\right)^\alpha$ is estimated by utilizing the bias-corrected estimator $\widehat{\mathcal{M}^u_\alpha}$ for each term of $p\left(\boldsymbol{p}\left[i\right]\right)$.

2.2.2. Lower Bounds

In the second part of this section, we exhibit lower bounds [7] on the task of estimating Rényi entropy to within an additive error. The techniques utilized for the task resemble to the ones used for establishing a lower bound on estimating Shannon entropy. Unsurprisingly, since Rényi entropy is a symmetric property, the fingerprint of a sample is satisfactory for the analysis, as described in Section 2.1. Acharya *et al.* construct two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ such that the difference, $|H_\alpha(\boldsymbol{p}) - H_\alpha(\boldsymbol{q})|$, is sufficiently big to distinguish $\boldsymbol{p}$ from $\boldsymbol{q}$. However, the fingerprint distributions, $\boldsymbol{p}_{\boldsymbol{\mathcal{F}}}$, $\boldsymbol{q}_{\boldsymbol{\mathcal{F}}}$, corresponding to $\boldsymbol{p}, \boldsymbol{q}$, respectively, have small total variation distance so that given a fingerprint it is impossible to decide which distribution it is obtained from when the number of samples is less than certain quantity. The following sets a bound on the total variation distance between the fingerprint distributions.

Theorem 2.41. ([7], Theorem 13) *Given distributions* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *such that* $\max_x \max\{\boldsymbol{p}_x, \boldsymbol{q}_x\} \leq \frac{\epsilon}{40m}$, *for Poisson sampling with* $M \sim \mathrm{Pois}(m)$, *it holds that*

$$\|\boldsymbol{p}_{\boldsymbol{\mathcal{F}}} - \boldsymbol{q}_{\boldsymbol{\mathcal{F}}}\| \leq \frac{\epsilon}{2} + 5 \sum_{\alpha} m^\alpha |\mathcal{M}_\alpha(\boldsymbol{p}) - \mathcal{M}_\alpha(\boldsymbol{q})|.$$

Thus, it is sufficient to build $\boldsymbol{p}, \boldsymbol{q}$ with distant Rényi entropies, yet having identical moments. For a probability distribution $\boldsymbol{p}$, let $\|\boldsymbol{p}\|_r = \left(\sum_{i=1}^{n} \left|\boldsymbol{p}[i]\right|^r\right)^{1/r}$, where r is a positive real number. We present the main ingredients of constructing such $\boldsymbol{p}, \boldsymbol{q}$ pairs.

Lemma 2.42. ([7], Lemma 14) *For every* $d \in \mathbb{N}$ *and noninteger* α, *there exist positive vectors* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ *such that*

$$\begin{aligned} &\|\boldsymbol{x}\|_r = \|\boldsymbol{y}\|_r, \ \ 1 \leq r \leq d-1 \\ &\|\boldsymbol{x}\|_d \neq \|\boldsymbol{y}\|_d, \\ &\|\boldsymbol{x}\|_\alpha \neq \|\boldsymbol{y}\|_\alpha. \end{aligned}$$

Definition 2.43. *For every positive integer* d *and every vector* $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, *construct a distribution* $\boldsymbol{p}^{\boldsymbol{x}}$ *with support size* dn *as follows*

$$\boldsymbol{p}^{\boldsymbol{x}}_{\boldsymbol{ij}} = \frac{|x_i|}{n\|\boldsymbol{x}\|_1}, \ \ 1 \leq i \leq d, \ 1 \leq j \leq n.^5$$

[5] Analyzing probability distributions with support size dn instead of the ones with support size n is only for practical reasons and does not affect the lower bound results stated above.

It only remains to construct distributions $\boldsymbol{p}, \boldsymbol{q}$ based on Lemma 2.42 and Definition 2.43 for three different cases of α and apply Theorem 2.41 to finalize the proof. Before stating the lower bound results on estimating Rényi entropy to within additive error, we give one last definition.

Definition 2.44. *Let $f(n) = \tilde{\tilde{\Omega}}\left(n^{\beta}\right)$ indicate that for all sufficiently large n and for all $\eta > 0$, $f(n) > n^{\beta-\eta}$ where $f : \mathbb{R} \to \mathbb{R}$ and $\beta \in \mathbb{R}$. That is, $f(n)$ grows polynomially in n with exponent not less than β.*

Theorem 2.45. ([7], Theorem 15) *For any integer $\alpha > 1$, $\Omega\left(n^{1-1/\alpha}\right)$ samples are necessary to estimate (with high probability) the Rényi entropy $H_{\alpha}(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

Theorem 2.46. ([7], Theorem 16) *For any noninteger $\alpha > 1$, $\tilde{\tilde{\Omega}}\left(n\right)$ samples are necessary to estimate (with high probability) the Rényi entropy $H_{\alpha}(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

Theorem 2.47. ([7], Theorem 17) *For any $1 > \alpha > 0$, $\tilde{\tilde{\Omega}}\left(n^{1/\alpha}\right)$ samples are necessary to estimate (with high probability) the Rényi entropy $H_{\alpha}(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

3. SAMP+PMF MODEL

We formally define the "unconventional" model.

Definition 3.1. *Let $\boldsymbol{p}$ be a probability distribution on $[n]$ and PMF denote a type of query which takes input $i \in [n]$ and returns its probability mass function $\boldsymbol{p}[i]$. The SAMP+PMF model is a model of query complexity in which both SAMP and PMF queries are utilized to interact with $\boldsymbol{p}$.*

Note that a trivial upper bound on estimating entropy is n, since one can learn $\boldsymbol{p}$ fully via n PMF queries. We start with the results regarding multiplicatively estimating entropy. As shown in [9, 12], estimating $H(\boldsymbol{p})$ (with high probability) to within a multiplicative factor of $1+\gamma$ requires between

$$\Omega\left(\frac{\log n}{\max(\gamma, \gamma^2)}\right) \cdot \frac{1}{H(\boldsymbol{p})} \tag{3.1}$$

and

$$O\left(\frac{\log n}{\gamma^2}\right) \cdot \frac{1}{H(\boldsymbol{p})} \tag{3.2}$$

SAMP+PMF queries. One disadvantage of these bounds is them depending quantitatively on the entropy $H(\boldsymbol{p})$ itself.

The task of estimating (with high probability) Shannon entropy to within additive error is examined in [10]. Canonne and Rubinfeld construct an algorithm with $O\left(\log^2 n\right)$ query complexity, demonstrating superiority of SAMP+PMF model to SAMP model in which $\Omega\left(\frac{n}{\log n}\right)$ samples are necessary for the same task. They build the algorithm on the following:

$$H(\boldsymbol{p}) = \sum_{i \in [n]} \boldsymbol{p}[i] \log \frac{1}{\boldsymbol{p}[i]} = \mathop{\mathbf{E}}_{i \sim \boldsymbol{p}}\left[\log \frac{1}{\boldsymbol{p}[i]}\right] \tag{3.3}$$

Since it is difficult to give an upper bound on the quantity $\log \frac{1}{\boldsymbol{p}[i]}$, Identity 3.3 is reconstructed. Note that the function $f\left(x\right) = x \log\left(\frac{1}{x}\right)$ is increasing for $x \in \left(0, \frac{1}{e}\right)$ and $\lim\limits_{x \to 0^+} f\left(x\right) = 0$. Then for any threshold $\tau \in \left(0, \frac{1}{e}\right)$,

$$\begin{aligned} H\left(\boldsymbol{p}\right) &= \sum_{i:\boldsymbol{p}[i] \geq \tau} \boldsymbol{p}\left[i\right] \log \frac{1}{\boldsymbol{p}\left[i\right]} + \sum_{i:\boldsymbol{p}[i] < \tau} \boldsymbol{p}\left[i\right] \log \frac{1}{\boldsymbol{p}\left[i\right]} \Rightarrow \\ H\left(\boldsymbol{p}\right) &\geq \sum_{i:\boldsymbol{p}[i] \geq \tau} \boldsymbol{p}\left[i\right] \log \frac{1}{\boldsymbol{p}\left[i\right]} = H\left(\boldsymbol{p}\right) - \sum_{i:\boldsymbol{p}[i] < \tau} \boldsymbol{p}\left[i\right] \log \frac{1}{\boldsymbol{p}\left[i\right]} \geq H\left(\boldsymbol{p}\right) - n \cdot \tau \log \frac{1}{\tau} \end{aligned} \tag{3.4}$$

Assume $\frac{\Delta}{n} < \frac{1}{2}$, and let $\tau = \frac{\frac{\Delta}{n}}{10 \log \frac{n}{\Delta}}$, so that $n \cdot \tau \log \frac{1}{\tau} \leq \frac{\Delta}{2}$. Define a function $\varphi\left(x\right) = \log \frac{1}{x} \mathbb{1}_{\{x \geq \tau\}}$. Then, Equation 3.4 implies that

$$H\left(\boldsymbol{p}\right) \geq \mathop{\mathbf{E}}_{i \sim \boldsymbol{p}}\left[\varphi\left(\boldsymbol{p}\left[i\right]\right)\right] \geq H\left(\boldsymbol{p}\right) - \frac{\Delta}{2}.$$

Consequently, estimating $\mathbf{E}_{i \sim \boldsymbol{p}}\left[\varphi\left(\boldsymbol{p}\left[i\right]\right)\right]$ to within additive $\frac{\Delta}{2}$ is sufficient for estimating $H\left(\boldsymbol{p}\right)$ to within additive Δ. Observe that $0 \leq \varphi\left(\boldsymbol{p}\left[i\right]\right) \leq \log \frac{1}{\tau} \approx \log \frac{n}{\Delta}$. Let $X_1, \ldots, X_m$ be independent samples drawn from $\boldsymbol{p}$ where $m = O\left(\frac{\log^2 \frac{n}{\Delta}}{\Delta^2}\right)$. If we compute the quantities $\varphi\left(\boldsymbol{p}\left[X_j\right]\right)$ via PMF queries and apply the Hoeffding bound on random variables $Y_j = \dfrac{\varphi\left(\boldsymbol{p}\left[X_j\right]\right)}{\log \frac{1}{\tau}}$ for $1 \leq j \leq m$,

$$\Pr\left[\left|\frac{1}{m}\sum_{j=1}^{m} \varphi\left(\boldsymbol{p}\left[X_j\right]\right) - \mathop{\mathbf{E}}_{i \sim \boldsymbol{p}}\left[\varphi\left(\boldsymbol{p}\left[i\right]\right)\right]\right| \geq \frac{\Delta}{2}\right] \leq 2e^{-\frac{\Delta^2 m}{\log^2 \frac{1}{\tau}}} \leq \frac{1}{3}.$$

Theorem 3.2. ([10], Theorem 10) *In the SAMP+PMF model, there exists an algorithm estimating Shannon entropy (with high probability) to within $\pm\Delta$ with sample complexity $O\left(\frac{\log^2 \frac{n}{\Delta}}{\Delta^2}\right)$.*

Apart from achieving an exponentially better bound, the simplicity of the algorithm compared to the entangled nature of the estimators described in Section 2.1 is sufficient to illuminate the power a PMF query adds to the model. Yet, how helpful is it for estimating Rényi entropy? We will examine this question in Section 4.3.

```
Algorithm: ESTIMATOR V
  • Fix $\tau = \frac{\frac{\Delta}{n}}{10 \log \frac{n}{\Delta}}$ and $m = \left\lceil \frac{\ln 6}{\Delta^2} \log^2 \frac{1}{\tau} \right\rceil$.
  • Draw $m$ independent samples $X_1, \ldots, X_m$ from $\boldsymbol{p}$.
  • Compute $Y_j = \log \frac{1}{\boldsymbol{p}[X_j]} \mathbb{1}_{\{\boldsymbol{p}[X_j] \geq \tau\}}$ by evaluating PMF on $X_j$ for $1 \leq j \leq m$.
  • Output $\frac{1}{m} \sum_{j=1}^{m} Y_j$.
```

Figure 3.1. SAMP+PMF Estimator of Shannon Entropy

We can extend the SAMP+PMF model by making a slight modification on a PMF query.

Definition 3.3. *Let $\boldsymbol{p}$ be a probability distribution on $[n]$ and CDF denote a type of query which takes input $i \in [n]$ and returns its cumulative distribution function $\sum_{j=1}^{i} \boldsymbol{p}[j]$. The SAMP+CDF model is a model in which both SAMP and CDF queries are utilized to interact with $\boldsymbol{p}$.*

Observe that since $\mathsf{PMF}(i) = \mathsf{CDF}(i) - \mathsf{CDF}(i-1)$, a PMF query is simulated by at most two CDF queries. Canonne and Rubinfeld [10] show that in certain cases the SAMP+CDF model is more powerful than the SAMP+PMF model in estimating Shannon entropy. More specifically, if a probability distribution $\boldsymbol{p}$ is known to be monotone, $\Omega(\log n)$ queries are necessary for estimating $H(\boldsymbol{p})$ in the SAMP+PMF model, however, there exists an algorithm in the SAMP+CDF model using $O\left(\log^2 \log n\right)$ queries for the same task. Does this hold true for estimating entropy of an arbitrary distribution? Answers to these and more questions are presented in the next chapter.

4. OPTIMAL BOUNDS FOR ESTIMATING ENTROPY WITH PMF QUERIES

4.1. Our Results, and Comparison with Prior Work

As described in Chapter 3, Canonne and Rubinfeld [10] build a SAMP+PMF algorithm estimating with high probability Shannon entropy to within ±1 using $O(\log^2 n)$ queries. In addition, they prove that $\Omega(\log n)$ queries are necessary for the task. Our first main result is an improved, optimal lower bound:

First main theorem. *In the SAMP+PMF model, $\Omega(\log^2 n)$ queries are necessary to estimate (with high probability) the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within ± 1.*

Remark 4.1. *Our lower bound and the lower bound for multiplicative estimation of Shannon entropy given in Expression 3.1 hold even under the promise that $H(\boldsymbol{p}) = \Theta(\log n)$. Note that Expression 3.1 yields a nonoptimal $\Omega(\log n)$ lower bound for additive estimation problem by taking $\gamma = \frac{1}{\log n}$.*

More precisely, Canonne and Rubinfeld show that $O(\frac{\log^2 n}{\Delta^2})$ queries are sufficient for estimating Shannon entropy to within $\pm\Delta$.[6] However, this result is trivial once $\Delta \leq \frac{\log n}{\sqrt{n}}$ since $\boldsymbol{p}$ can be learned exactly via n PMF queries. In fact, our first main theorem gives a matching lower bound for essentially the full range of Δ: we prove that $\Omega(\frac{\log^2 n}{\Delta^2})$ SAMP+PMF queries are necessary for any $\frac{1}{n^{0.4999}} \leq \Delta \leq \frac{\log n}{16 \cdot 10^6}$.

Our second main result is regarding the estimation of Rényi entropy $H_\alpha(\boldsymbol{p})$ for various parameters $0 \leq \alpha \leq \infty$. Recall that Acharya *et al.* [7] establish three different lower and upper bound pairs on additively estimating $H_\alpha(\boldsymbol{p})$ in the SAMP model. They prove that $\Theta(n^{1-1/\alpha})$ samples are necessary and sufficient when $\alpha > 1$ is an integer;

[6] They actually state $O(\frac{\log^2(n/\Delta)}{\Delta^2})$, but this is the same as $O(\frac{\log^2 n}{\Delta^2})$ because the range of interest is $\frac{\log n}{\sqrt{n}} \leq \Delta \leq \log n$.

$\tilde{\tilde{\Omega}}\left(n\right)$ samples are necessary when $\alpha > 1$ is a noninteger; $\tilde{\tilde{\Omega}}\left(n^{1/\alpha}\right)$ samples are necessary when $1 > \alpha > 0$. We give matching upper and lower bounds on estimating $H_\alpha(\boldsymbol{p})$ for all $\alpha > 1$. Apparently, PMF queries provide no advantage in estimating Rényi entropy for integer α, whereas they *are* advantageous for noninteger α.

Second main theorem. *Let $\alpha > 1$ be a real number. In the SAMP+PMF model, $\Omega\left(\frac{n^{1-1/\alpha}}{2^{2\Delta}}\right)$ queries are necessary and $O\left(\frac{n^{1-1/\alpha}}{\left(1-2^{(1-\alpha)\Delta}\right)^2}\right)$ queries are sufficient to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

We show that our two bounds extend to the SAMP+CDF model, thus, answering the question of the previous chapter. In addition, we give a matching lower bound for estimating support size to within $\pm\epsilon n$ in the SAMP+CDF model. Lastly, we provide an upper bound on additively estimating Tsallis entropy in the SAMP+PMF model.

4.2. First Main Theorem

We present a well-known fact which is of key importance in establishing a lower bound on additively estimating Shannon entropy in the SAMP+PMF model.

Lemma 4.2. *For $\lambda \in \left(0, \frac{1}{4}\right]$, $\Theta\left(\frac{1}{\lambda^2}\right)$ samples are necessary and sufficient to distinguish between the uniform distribution $\boldsymbol{p_1} = \left(\frac{1}{2}, \frac{1}{2}\right)$ and the biased distribution $\boldsymbol{p_2} = \left(\frac{1}{2} + \lambda, \frac{1}{2} - \lambda\right)$.*[7]

Proof. We start with proving the upper bound. Let $X_1, \ldots, X_m$ be m independent samples drawn from $\boldsymbol{p}$ which is a probability distribution promised to be either a uniform or a biased distribution. Then $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{X_i=1\}}$ is an unbiased estimator for $\boldsymbol{p}[1]$, where $\mathbb{1}_{\{E\}}$ is an indicator function of an event E. We need to approximate (with high probability) $\boldsymbol{p}[1]$ to within some $\epsilon < \frac{\lambda}{2}$ to properly distinguish between two

[7] The range of λ can be easily extended to $\left(0, \frac{1}{2} - \xi\right]$, where ξ is an arbitrarily small constant.

distributions. By applying the Hoeffding bound,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\{X_i=1\}}-\boldsymbol{p}\left[1\right]\right|>\epsilon\right]\leq 2e^{-2\epsilon^2 m}.$$

Obviously, to bound the right-hand side of the inequality $m = O\left(\frac{1}{\lambda^2}\right)$ samples are sufficient. Now we prove the lower bound.

Definition 4.3. *Let $\boldsymbol{d_1}, \boldsymbol{d_2}$ be discrete probability distributions. The Kullback-Leibler (KL) divergence is defined as*

$$\mathbf{KL}\left(\boldsymbol{d_1}\|\boldsymbol{d_2}\right)=\sum_{x}\boldsymbol{d_1}\left[x\right]\log\frac{\boldsymbol{d_1}\left[x\right]}{\boldsymbol{d_2}\left[x\right]}. \tag{4.1}$$

By convention, $\boldsymbol{d_1}\left[x\right]\log\boldsymbol{d_1}\left[x\right]$ is set to 0 if $\boldsymbol{d_1}\left[x\right]=0$. We exploit the KL-divergence for the particular class of probability distributions.

Fact 4.4. *Let $X_1, \dots, X_m$ be m i.i.d random variables drawn from $\boldsymbol{d_1}$ (respectively $\boldsymbol{d_2}$), where $m \in \mathbb{N}$. Let $\boldsymbol{d_1^m}$ (respectively $\boldsymbol{d_2^m}$) denote a joint probability distribution of random variables $X_1, \dots, X_m$. Then*

$$\mathbf{KL}\left(\boldsymbol{d_1^m}\|\boldsymbol{d_2^m}\right)=m\,\mathbf{KL}\left(\boldsymbol{d_1}\|\boldsymbol{d_2}\right).$$

Proof. By definition $\boldsymbol{d_1^m}=\left(\boldsymbol{d_1}\right)^m$ and $\boldsymbol{d_2^m}=\left(\boldsymbol{d_2}\right)^m$. We prove the statement via strong induction.

Base case: For $m = 2$,

$$\begin{aligned}\mathbf{KL}\left(\boldsymbol{d_1^2}\|\boldsymbol{d_2^2}\right)&=\sum_{X_1}\sum_{X_2}\boldsymbol{d_1}\left[X_1\right]\boldsymbol{d_1}\left[X_2\right]\cdot\log\left(\frac{\boldsymbol{d_1}\left[X_1\right]\boldsymbol{d_1}\left[X_2\right]}{\boldsymbol{d_2}\left[X_1\right]\boldsymbol{d_2}\left[X_2\right]}\right)\\&=\sum_{X_1}\sum_{X_2}\boldsymbol{d_1}\left[X_1\right]\boldsymbol{d_1}\left[X_2\right]\cdot\left(\log\frac{\boldsymbol{d_1}\left[X_1\right]}{\boldsymbol{d_2}\left[X_1\right]}+\log\frac{\boldsymbol{d_1}\left[X_2\right]}{\boldsymbol{d_2}\left[X_2\right]}\right)\\&=\sum_{X_1}\boldsymbol{d_1}\left[X_1\right]\log\frac{\boldsymbol{d_1}\left[X_1\right]}{\boldsymbol{d_2}\left[X_1\right]}\sum_{X_2}\boldsymbol{d_1}\left[X_2\right]+\sum_{X_1}\boldsymbol{d_1}\left[X_1\right]\sum_{X_2}\boldsymbol{d_1}\left[X_2\right]\log\frac{\boldsymbol{d_1}\left[X_2\right]}{\boldsymbol{d_2}\left[X_2\right]}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{X_1} \boldsymbol{d_1}[X_1] \log \frac{\boldsymbol{d_1}[X_1]}{\boldsymbol{d_2}[X_1]} + \sum_{X_2} \boldsymbol{d_1}[X_2] \log \frac{\boldsymbol{d_1}[X_2]}{\boldsymbol{d_2}[X_2]} \\
&= 2\,\mathbf{KL}\left(\boldsymbol{d_1} \| \boldsymbol{d_2}\right).
\end{aligned}
$$

Induction step: Assume that the statement is true for all $i \leq m-1$. Then

$$
\begin{aligned}
\mathbf{KL}\left(\boldsymbol{d_1^m} \| \boldsymbol{d_2^m}\right) &= \sum_{X_1} \cdots \sum_{X_m} \left(\prod_{i=1}^{m} \boldsymbol{d_1}[X_i] \log \left(\prod_{j=1}^{m} \frac{\boldsymbol{d_1}[X_j]}{\boldsymbol{d_2}[X_j]} \right) \right) \\
&= \sum_{X_1} \cdots \sum_{X_m} \left(\prod_{i=1}^{m} \boldsymbol{d_1}[X_i] \left(\sum_{j=1}^{m} \log \frac{\boldsymbol{d_1}[X_j]}{\boldsymbol{d_2}[X_j]} \right) \right) \\
&= \sum_{X_1} \boldsymbol{d_1}[X_1] \log \frac{\boldsymbol{d_1}[X_1]}{\boldsymbol{d_2}[X_1]} \sum_{X_2} \boldsymbol{d_1}[X_2] \cdots \sum_{X_m} \boldsymbol{d_1}[X_m] \\
&\quad + \sum_{X_1} \boldsymbol{d_1}[X_1] \sum_{X_2} \cdots \sum_{X_m} \left(\prod_{i=2}^{m} \boldsymbol{d_1}[X_i] \left(\sum_{j=2}^{m} \log \frac{\boldsymbol{d_1}[X_j]}{\boldsymbol{d_2}[X_j]} \right) \right) \\
&= \sum_{X_1} \boldsymbol{d_1}[X_1] \log \frac{\boldsymbol{d_1}[X_1]}{\boldsymbol{d_2}[X_1]} + \sum_{X_2} \cdots \sum_{X_m} \left(\prod_{i=2}^{m} \boldsymbol{d_1}[X_i] \left(\sum_{j=2}^{m} \log \frac{\boldsymbol{d_1}[X_j]}{\boldsymbol{d_2}[X_j]} \right) \right) \\
&= \mathbf{KL}\left(\boldsymbol{d_1} \| \boldsymbol{d_2}\right) + \mathbf{KL}\left(\boldsymbol{d_1^{m-1}} \| \boldsymbol{d_2^{m-1}}\right) \\
&= m\,\mathbf{KL}\left(\boldsymbol{d_1} \| \boldsymbol{d_2}\right).
\end{aligned}
$$

□

Lemma 4.5. *Let $\widehat{\boldsymbol{d_1}}, \widehat{\boldsymbol{d_2}}$ be discrete probability distributions on the universe U, and let $f : U \to [0, C]$ be a real-valued function for some positive constant C. Then*

$$
\left| \underset{\widehat{\boldsymbol{d_1}}}{\mathbf{E}}[f(x)] - \underset{\widehat{\boldsymbol{d_2}}}{\mathbf{E}}[f(x)] \right| \leq C \cdot \|\widehat{\boldsymbol{d_1}} - \widehat{\boldsymbol{d_2}}\|_1.
$$

Proof.

$$
\begin{aligned}
\left| \underset{\widehat{\boldsymbol{d_1}}}{\mathbf{E}}[f(x)] - \underset{\widehat{\boldsymbol{d_2}}}{\mathbf{E}}[f(x)] \right| &= \left| \sum_{x \in U} \widehat{\boldsymbol{d_1}}[x] f(x) - \sum_{x \in U} \widehat{\boldsymbol{d_2}}[x] f(x) \right| \\
&= \left| \sum_{x \in U} \left(\widehat{\boldsymbol{d_1}}[x] - \widehat{\boldsymbol{d_2}}[x] \right) f(x) \right| \\
&\leq \sum_{x \in U} \left| \left(\widehat{\boldsymbol{d_1}}[x] - \widehat{\boldsymbol{d_2}}[x] \right) \right| f(x) \\
&\leq C \sum_{x \in U} \left| \left(\widehat{\boldsymbol{d_1}}[x] - \widehat{\boldsymbol{d_2}}[x] \right) \right|
\end{aligned}
$$

$$= C \cdot \|\widehat{\boldsymbol{d_1}} - \widehat{\boldsymbol{d_2}}\|_1. \qquad \square$$

As a final ingredient we give an upper bound on the KL-divergence between the biased distribution and the uniform distribution.

$$\begin{aligned}
\mathbf{KL}\left(\boldsymbol{p_2} \| \boldsymbol{p_1}\right) &= \left(\frac{1}{2} + \lambda\right) \log \frac{\frac{1}{2} + \lambda}{\frac{1}{2}} + \left(\frac{1}{2} - \lambda\right) \log \frac{\frac{1}{2} - \lambda}{\frac{1}{2}} \\
&= \frac{1}{2} \log\left(\left(1 - 2\lambda\right)\left(1 + 2\lambda\right)\right) + \lambda \log\left(\frac{1 + 2\lambda}{1 - 2\lambda}\right) \\
&\leq \lambda \log\left(\frac{1 + 2\lambda}{1 - 2\lambda}\right) \\
&= \frac{\lambda}{\ln 2} \ln\left(1 + \frac{4\lambda}{1 - 2\lambda}\right) \\
&\leq \frac{4\lambda^2}{\ln 2} \cdot \frac{1}{1 - 2\lambda} \\
&\leq \frac{8\lambda^2}{\ln 2},
\end{aligned} \tag{4.2}$$

where the second inequality follows from the exponential inequality.

Suppose for the sake of contradiction that there exists a hypothetical algorithm $\mathcal{D}$ such that given $m = o\left(\frac{1}{\lambda^2}\right)$ independent samples, decides whether $\boldsymbol{p} = \boldsymbol{p_1}$ or $\boldsymbol{p} = \boldsymbol{p_2}$ with probability of error at most $\frac{1}{3}$. Let $\mathcal{D}$ output 1 to indicate that $\boldsymbol{p} = \boldsymbol{p_1}$ and output 0 to indicate that $\boldsymbol{p} = \boldsymbol{p_2}$. Denote by $\mathcal{X}$ the set of m independent samples $\{X_1, \ldots X_m\}$. Then

$$\Pr\left[\mathcal{D}\left(\mathcal{X} \sim \boldsymbol{p_1^m}\right) = 1\right] \geq \frac{2}{3} \quad \text{and} \quad \Pr\left[\mathcal{D}\left(\mathcal{X} \sim \boldsymbol{p_2^m}\right) = 0\right] \geq \frac{2}{3}.$$

In other words,

$$\mathop{\mathbf{E}}_{\mathcal{X} \sim \boldsymbol{p_1^m}}\left[\mathcal{D}\left(\mathcal{X}\right)\right] \geq \frac{2}{3} \quad \text{and} \quad \mathop{\mathbf{E}}_{\mathcal{X} \sim \boldsymbol{p_2^m}}\left[\mathcal{D}\left(\mathcal{X}\right)\right] \leq \frac{1}{3},$$

which gives

$$\underset{\mathcal{X}\sim\boldsymbol{p}_{\mathbf{1}}^{\boldsymbol{m}}}{\mathbf{E}}\left[\mathcal{D}\left(\mathcal{X}\right)\right]-\underset{\mathcal{X}\sim\boldsymbol{p}_{\mathbf{2}}^{\boldsymbol{m}}}{\mathbf{E}}\left[\mathcal{D}\left(\mathcal{X}\right)\right]\geq\frac{1}{3}. \tag{4.3}$$

Letting $\widehat{\boldsymbol{d}_{\mathbf{1}}}=\boldsymbol{p}_{\mathbf{1}}^{\boldsymbol{m}}$, $\widehat{\boldsymbol{d}_{\mathbf{2}}}=\boldsymbol{p}_{\mathbf{2}}^{\boldsymbol{m}}$ and $f=\mathcal{D}$, Lemma 4.5 and Inequality 4.3 imply

$$\|\boldsymbol{p}_{\mathbf{1}}^{\boldsymbol{m}}-\boldsymbol{p}_{\mathbf{2}}^{\boldsymbol{m}}\|_1\geq\frac{1}{3}. \tag{4.4}$$

Then

$$\begin{aligned}
o\left(\frac{1}{\lambda^2}\right)=m&=\frac{\mathbf{KL}\left(\boldsymbol{p}_{\mathbf{1}}^{\boldsymbol{m}}\|\boldsymbol{p}_{\mathbf{2}}^{\boldsymbol{m}}\right)}{\mathbf{KL}\left(\boldsymbol{p}_{\mathbf{1}}\|\boldsymbol{p}_{\mathbf{2}}\right)} && \text{(Fact 4.4)}\\
&\geq\frac{1}{2\ln 2}\cdot\|\boldsymbol{p}_{\mathbf{1}}^{\boldsymbol{m}}-\boldsymbol{p}_{\mathbf{2}}^{\boldsymbol{m}}\|_1^2\cdot\frac{1}{\mathbf{KL}\left(\boldsymbol{p}_{\mathbf{1}}\|\boldsymbol{p}_{\mathbf{2}}\right)} && \text{(Pinsker's inequality)}\\
&\geq\frac{1}{18\ln 2}\cdot\frac{1}{\mathbf{KL}\left(\boldsymbol{p}_{\mathbf{1}}\|\boldsymbol{p}_{\mathbf{2}}\right)} && \text{(Inequality 4.4)}\\
&\geq\frac{1}{144\lambda^2}=\Omega\left(\frac{1}{\lambda^2}\right), && \text{(Inequality 4.2)}
\end{aligned}$$

which is a contradiction. □

We establish a tight lower bound on estimating Shannon entropy in the following theorem.

Theorem 4.6. *In the SAMP+PMF model, $\Omega\left(\frac{\log^2 n}{\Delta^2}\right)$ queries are necessary to estimate (with high probability) the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$, where $\frac{1}{n^{0.4999}}\leq\Delta\leq\frac{\log n}{16\cdot 10^6}$.*

Proof. We will show that a hypothetical SAMP+PMF algorithm $\mathcal{E}$ that can estimate the entropy of an unknown distribution on $[n]$ to within $\pm\Delta$ using $o\left(\frac{\log^2 n}{\Delta^2}\right)$ queries would contradict Lemma 4.2 stating that $\Omega(1/\lambda^2)$ coin tosses are necessary to determine whether a given coin is fair, or comes up heads with probability $1/2+\lambda$.

The idea is to use the given coin to realize the probability distribution that $\mathcal{E}$ will work on. Let n be the smallest one millionth power of a natural number that satisfies $\frac{4\cdot 10^6\Delta}{\log n} \leq \lambda$.[8] Partition the domain $[n]$ into $M = n^{0.999999}$ consecutive blocks $I_1, \ldots, I_M$, each containing $K = \frac{n}{M} = n^{0.000001}$ elements. Each block will be labeled either as a tails or a heads block. The internal distribution of each heads block is uniform, i.e. each element has probability mass $\frac{1}{MK} = \frac{1}{n}$. In each tails block, the first element has probability mass $\frac{1}{n^{0.999999}}$, while the rest of the elements have probability mass 0. Note that the total probability mass of each block is $K \cdot \frac{1}{MK} = \frac{1}{M} = \frac{1}{n^{0.999999}}$, regardless of its label. We will now describe a costly method of constructing a probability distribution $\boldsymbol{p}$ of this kind, using a coin that comes up heads with probability d:

- Throw the coin M times to obtain the outcomes $X_1, \ldots, X_M$,
- Set the label of block I_m to X_m, for all $m \in [M]$.

Let X be the number of heads blocks in $\boldsymbol{p}$. Then $\mu = \mathbf{E}[X] = Md$. Let $\overline{X} = \frac{X}{M}$ denote the proportion of heads blocks in $\boldsymbol{p}$. Then we can calculate the entropy $H(\boldsymbol{p})$ by calculating the individual entropies of the blocks. For a heads block, the entropy is $K \cdot \frac{1}{MK} \cdot \log(MK) = \frac{1}{M}\log n$. The entropy of a tails block is $\frac{1}{n^{0.999999}}\log(n^{0.999999}) = \frac{0.999999}{M}\log n$. Since there are $M\overline{X}$ heads blocks and $M(1-\overline{X})$ tails blocks, the total entropy becomes

$$\begin{aligned} H(\boldsymbol{p}) &= M\overline{X} \cdot \tfrac{1}{M}\log n + M(1-\overline{X}) \cdot \tfrac{0.999999}{M}\log n \\ &= \overline{X}\log n + 0.999999(1-\overline{X})\log n \qquad (4.5) \\ &= (0.999999 + 0.000001\overline{X})\log n. \end{aligned}$$

Note that this function is monotone with respect to $\overline{X}$. Define two families of distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ constructed by the above process, taking d to be $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{2} + \lambda$, respectively. Let $\boldsymbol{p_1}$ (respectively $\boldsymbol{p_2}$) be a probability distribution randomly chosen from $\mathcal{P}_1$ (respectively $\mathcal{P}_2$).

[8] Since Lemma 4.2 holds for $\lambda \in \left(0, \frac{1}{2} - \xi\right]$, where ξ is an arbitrarily small constant, the upper bound on Δ can be extended to $\frac{(1-2\xi)\log n}{8\cdot 10^6}$.

Proposition 4.7. *$\boldsymbol{p_1}$ has entropy at most $0.9999995\log n + \Delta$ with high probability.*

Proof. We prove this by using the Chernoff bound on the number of heads blocks in the distribution.

$$\begin{aligned}
\Pr\Big[X \geq \big(p_1 + \tfrac{10^6\Delta}{\log n}\big)M\Big] &= \Pr\Big[X \geq \tfrac{M}{2}\big(1 + \tfrac{2\cdot 10^6\Delta}{\log n}\big)\Big] \\
&\leq \exp\Big(-\frac{\frac{4\cdot 10^{12}\Delta^2}{\log^2 n}}{2 + \frac{2\cdot 10^6\Delta}{\log n}}\frac{M}{2}\Big) \\
&= \exp\Big(-\frac{10^{12}\Delta^2 M}{\log n(\log n + 10^6\Delta)}\Big) \\
&\leq \exp\Big(-\frac{10^{12}\cdot n^{0.999999}/n^{0.999998}}{\log^2 n(1+10^6)}\Big) \\
&= \exp\Big(-\frac{10^{12}\cdot n^{0.000001}}{\log^2 n(1+10^6)}\Big) \\
&= o(1).
\end{aligned}$$

The last term indicates that the number of heads blocks $X < \big(p_1 + \frac{10^6\Delta}{\log n}\big)M$, and the proportion of the heads blocks $\overline{X} < \big(p_1 + \frac{10^6\Delta}{\log n}\big)$ with high probability. Thus, with high probability

$$H[\boldsymbol{p_1}] = (0.999999 + 0.000001\overline{X})\log n < 0.9999995\log n + \Delta. \qquad \square$$

Proposition 4.8. *$\boldsymbol{p_2}$ has entropy at least $0.9999995\log n + 3\Delta$ with high probability.*

Proof. We find a similar bound by;

$$\begin{aligned}
\Pr\Big[X \leq \big(p_2 - \tfrac{10^6\Delta}{\log n}\big)M\Big] &= \Pr\Big[X \leq p_2 M\big(1 - \tfrac{10^6\Delta}{p_2\log n}\big)\Big] \\
&\leq \exp\Big(-\frac{\frac{10^{12}\Delta^2}{p_2^2\log^2 n}}{2}p_2 M\Big) \\
&= \exp\Big(-\frac{10^{12}\Delta^2 M}{2p_2\log^2 n}\Big) \\
&= \exp\Big(-\frac{10^{12}\Delta^2 M}{2\left(\frac{1}{2}+\lambda\right)\log^2 n}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \exp\left(- \frac{n^{0.000001}}{\log^2 n} \right) \\
&= o(1).
\end{aligned}
$$

The last term indicates that the number of heads blocks $X > \big(p_2 - \frac{10^6\Delta}{\log n}\big)M$, and the proportion of the heads blocks $\overline{X} > \big(p_2 - \frac{10^6\Delta}{\log n}\big)$ with high probability. Thus, with high probability

$$
\begin{aligned}
H[\boldsymbol{p_2}] &= (0.999999 + 0.000001\overline{X}) \log n \\
&> \left(0.999999 + 0.000001\left(\frac{1}{2} + \lambda - \tfrac{10^6\Delta}{\log n}\right)\right) \log n \\
&= 0.9999995 \log n + 0.000001(\lambda - \tfrac{10^6\Delta}{\log n}) \log n \qquad (4.6) \\
&\geq 0.9999995 \log n + 0.000001(\tfrac{3\cdot 10^6\Delta}{\log n}) \log n \\
&= 0.9999995 \log n + 3\Delta. \qquad \square
\end{aligned}
$$

Since the entropies of $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ are sufficiently far apart from each other, our hypothetical estimator $\mathcal{E}$ can be used to determine whether the underlying coin has probability p_1 or p_2 associated with it. To arrive at the contradiction we want, we must ensure that the coin is not thrown too many times during this process. This is achieved by constructing the distribution "on-the-fly" [10] during the execution of $\mathcal{E}$, throwing the coin only when it is required to determine the label of a previously undefined block:

When $\mathcal{E}$ makes a SAMP query, we choose a block I_m uniformly at random (since each block has probability mass $\frac{1}{M}$), and then flip the coin for I_m to decide its label if it is yet undetermined. We then draw a sample $i \sim \mathbf{d_m}$ from I_m, where $\mathbf{d_m}$ is the normalized distribution of the m^{th} block.

When $\mathcal{E}$ makes a PMF query on $i \in [n]$, we flip the coin to determine the label of the associated block I_m if it is yet undetermined. We then return the probability mass of i.

By this procedure, the queries of $\mathcal{E}$ about the probability distribution $\boldsymbol{p}$ (known to be either $\boldsymbol{p_1}$ or $\boldsymbol{p_2}$) can be answered by using at most one coin flip per query, i.e. $o\left(\frac{\log^2 n}{\Delta^2}\right)$ times in total.

Since we selected n so that $1/\lambda^2 = \Theta(\frac{\log^2 n}{\Delta^2})$, this would mean that it is possible to distinguish between the two coins using only $o(1/\lambda^2)$ throws, which is a contradiction, letting us conclude that no algorithm can estimate the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with high probability making $o\left(\frac{\log^2 n}{\Delta^2}\right)$ queries. □

We now give a similar lower bound for the SAMP+CDF model.

Corollary 4.9. *In the SAMP+CDF model, any algorithm estimating (with high probability) the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ must make $\Omega\left(\frac{\log^2 n}{\Delta^2}\right)$ queries.*

Proof. The construction is identical to the one in the proof of Theorem 4.6, except that we now have to describe how the CDF queries of the estimation algorithm must be answered using the coin:

When $\mathcal{E}$ makes a CDF query on $i \in [n]$, we flip the coin to determine the label of the associated block I_m if this is necessary. We then return the sum of the total probability mass of the blocks preceding I_m (which is $\frac{m-1}{M}$, since each block has a total probability mass of $\frac{1}{M}$ regardless of its label) and the probability masses of the elements from the beginning of I_m up to and including i itself. At most one coin flip per CDF query is therefore sufficient. □

4.3. Second Main Theorem

4.3.1. Lower Bound

We present another well-known fact about distinguishing coins.

Fact 4.10. *$\Omega\left(\frac{1}{\lambda}\right)$ samples are necessary to distinguish between the distribution $\boldsymbol{p_1} = (1,0)$ and the distribution $\boldsymbol{p_2} = (1-\lambda, \lambda)$.*

Proof. Let $X_1, \ldots, X_m$ be m independent samples drawn from $\boldsymbol{p}$, which is a probability distribution promised to be either $\boldsymbol{p_1}$ or $\boldsymbol{p_2}$. Note that observing the domain element 2 even once suffices to distinguish between two distributions, since $\boldsymbol{p_1}[2] = 0$. The probability of the most unfortunate case, not observing any 2 when $\boldsymbol{p} = \boldsymbol{p_2}$, is

$$\Pr\left[\sum_{i=1}^{m} \mathbb{1}_{\{X_i=1\}} = m\right] = (1-\lambda)^m \leq e^{-\lambda m} \tag{4.7}$$

where the inequality follows from the exponential inequality. Obviously, to bound the right-hand side of Inequality 4.7, $m = \Omega\left(\frac{1}{\lambda}\right)$ samples are necessary. □

We establish a lower bound on estimating Rényi entropy in the following theorem.

Theorem 4.11. *For any $\alpha > 1$, $\Omega\left(\frac{n^{1-1/\alpha}}{2^{2\Delta}}\right)$ SAMP+PMF queries are necessary to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

Proof. We will first prove the theorem for rational α, and show that it remains valid for irrationals at the end.

The proof has the same structure as that of Theorem 4.6. One difference is that we reduce from the problem of distinguishing a maximally biased coin that never comes up tails from a less biased one (instead of the problem of distinguishing a fair coin from a biased one).

Suppose that we are given a coin whose probability of coming up heads is promised to be either $p_1 = 1$ or $p_2 = 1-\lambda$ for a specified number λ, and we must determine which is the case. As Fact 4.10 indicates, this task requires at least $\Omega(1/\lambda)$ coin throws. We will show that this fact is contradicted if one assumes that there exist natural numbers s and t, where $\alpha = \frac{s}{t} > 1$, such that it is possible to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ using an algorithm, say $\mathcal{R}$, that makes only $o(\frac{n^{1-1/\alpha}}{2^{2\Delta}})$ SAMP+PMF queries.

Let n be the smallest number of the form $\left(\lceil 2^{2\Delta}\rceil j\right)^s$ that satisfies $\frac{5\cdot\lceil 2^{2\Delta}\rceil}{n^{1-1/\alpha}} \leq \lambda$, where j is some natural number. Partition $[n]$ into $M = \frac{n^{1-1/\alpha}}{\lceil 2^{2\Delta}\rceil}$ consecutive blocks $I_1, I_2, \ldots I_M$, each of size $K = \lceil 2^{2\Delta}\rceil \cdot n^{1/\alpha}$. As in the proof of Theorem 4.6, a probability distribution $\boldsymbol{p}$ can be realized by throwing a given coin M times to obtain the outcomes $X_1, \ldots, X_M$, and setting the label of block I_m to X_m, for all $m \in [M]$, where each member of each heads block again has probability mass $1/n$. The first member of each tails block has probability mass $\frac{\lceil 2^{2\Delta}\rceil}{n^{1-1/\alpha}}$, and the remaining members have probability mass 0. We again have that each block has total probability mass $\frac{K}{n} = \frac{\lceil 2^{2\Delta}\rceil n^{1/\alpha}}{n} = \frac{1}{M}$ regardless of its label, so this process always results in a legal probability distribution.

If the coin is maximally biased, then $\boldsymbol{p}$ becomes the uniform distribution, and $H_\alpha(\boldsymbol{p}) = \log n$. We will examine the probability of the same distribution being obtained using the less biased coin. Let $\mathcal{P}_2$ be the family of distributions constructed by the process described above, using a coin with probability p_2 of coming up heads. Let $\boldsymbol{p_2}$ be a probability distribution randomly chosen from $\mathcal{P}_2$. The probability of the undesired case where $\boldsymbol{p_2}$ is the uniform distribution is

$$\Pr\left[\boldsymbol{p_2} = \mathcal{U}\left([n]\right)\right] = p_2^M = (1-\lambda)^M \leq \left(1 - \frac{5\cdot\lceil 2^{2\Delta}\rceil}{n^{1-1/\alpha}}\right)^M \leq e^{-\frac{5\cdot\lceil 2^{2\Delta}\rceil}{n^{1-1/\alpha}}M} = e^{-5} \leq \frac{1}{1000} \quad .$$

That is, with probability ≥ 0.999, $\boldsymbol{p_2}$ has at least one element with probability mass $\frac{\lceil 2^{2\Delta}\rceil n^{\frac{1}{\alpha}}}{n}$. Let X be the number of heads outcomes, and let B and W denote the number of elements with probability mass $\frac{1}{n}$ and $\frac{\lceil 2^{2\Delta}\rceil n^{\frac{1}{\alpha}}}{n}$, respectively. It is not difficult to see that $B = K \cdot X$ and $W = M - X$. We just showed that $X < M$ with high probability.

Then the Rényi entropy of the constructed distribution $\boldsymbol{p_2} \in \mathcal{P}_2$ is, with high probability:

$$\begin{aligned}
H_\alpha(\boldsymbol{p_2}) &= \frac{1}{1-\alpha}\log\left(B\cdot\frac{1}{n^\alpha}+W\cdot\left(\frac{\lceil 2^{2\Delta}\rceil n^{\frac{1}{\alpha}}}{n}\right)^\alpha\right)\\
&= \frac{1}{1-\alpha}\log\left(\frac{K\cdot X/n+(M-X)\lceil 2^{2\Delta}\rceil^\alpha}{n^{\alpha-1}}\right)\\
&= \log n - \frac{1}{\alpha-1}\log\left(K\cdot X/n+(M-X)\lceil 2^{2\Delta}\rceil^\alpha\right)\\
&\le \log n - \frac{1}{\alpha-1}\log\left(\lceil 2^{2\Delta}\rceil^\alpha\right) < \log n - 2\Delta.
\end{aligned}$$

Because $H_\alpha(\mathcal{U}([n])) - H_\alpha(\boldsymbol{p_2}) > 2\Delta$, $\mathcal{R}$ has to be able to distinguish $\mathcal{U}([n])$ and $\boldsymbol{p_2}$ with high probability. We can then perform a simulation of $\mathcal{R}$ involving an "on-the-fly" construction of distribution $\boldsymbol{p}$ exactly as described in the proof of Theorem 4.6. As discussed in Section 4.2, this process requires no more coin throws than the number of SAMP+PMF queries made by $\mathcal{R}$, allowing us to determine the type of the coin using only $o(\frac{n^{1-1/\alpha}}{2^{2\Delta}})$, that is, $o(1/\lambda)$ tosses with high probability, a contradiction.

Having thus proven the statement for rational α, it is straightforward to cover the case of irrational α: Note that $H_\alpha(\boldsymbol{p})$ is a continuous function of α for fixed $\boldsymbol{p}$. Given any $\boldsymbol{p}$ and ε, for any irrational number α_i greater than 1, there exists a rational α_r which is so close to α_i such that $H_{\alpha_i}(\boldsymbol{p}) - H_{\alpha_r}(\boldsymbol{p}) < \varepsilon$. An efficient entropy estimation method for some irrational value of α would therefore imply the existence of an equally efficient method for some rational value, contradicting the result obtained above. □

These results are generalized to the SAMP+CDF model in the same way as in Section 4.2.

Corollary 4.12. *For any $\alpha > 1$, $\Omega\left(\frac{n^{1-1/\alpha}}{2^{2\Delta}}\right)$ SAMP+PMF or SAMP+CDF queries are necessary to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

4.3.2. Upper Bound

We now show that PMF queries are useful for estimating H_α for noninteger α.

Theorem 4.13. *For any number $\alpha > 1$, there exists an algorithm estimating (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with $O\left(\frac{n^{1-1/\alpha}}{\left(1-2^{(1-\alpha)\Delta}\right)^2}\right)$ SAMP+PMF queries.*

Proof. We will prove this statement for rational α. The generalization to irrational α discussed in the proof of Theorem 4.11 can be applied. Recall that Rényi entropy can be expressed as $H_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha}\log \mathcal{M}_\alpha(\boldsymbol{p})$, where $\mathcal{M}_\alpha(\boldsymbol{p}) = \sum_{i=1}^{n} (\boldsymbol{p}[i])^\alpha$ is the α^{th} moment of $\boldsymbol{p}$. Then, estimating $H_\alpha(\boldsymbol{p})$ to an additive accuracy of $\pm\Delta$ is equivalent to estimating $\mathcal{M}_\alpha(\boldsymbol{p})$ to a multiplicative accuracy of $2^{\pm\Delta(1-\alpha)}$. Therefore, we construct a multiplicative estimator for $\mathcal{M}_\alpha(\boldsymbol{p})$.

Let $\gamma = 1 - 2^{(1-\alpha)\Delta}$ and $m = \left\lceil \frac{100n^{1-1/\alpha}}{\gamma^2} \right\rceil$, and let $X_1, \ldots, X_m$ be i.i.d. random variables drawn from $\boldsymbol{p}$. Define $Y_i = (\boldsymbol{p}[X_i])^{\alpha-1}$, where $\boldsymbol{p}[X_i]$ can be calculated using a PMF query on X_i for $1 \leq i \leq m$. Note that

$$\mathbf{E}[Y_i] = \sum_{j=1}^{n} \boldsymbol{p}[j]\,(\boldsymbol{p}[j])^{\alpha-1} = \sum_{j=1}^{n} (\boldsymbol{p}[j])^{\alpha} = \mathcal{M}_\alpha(\boldsymbol{p}).$$

Then $\frac{1}{m}\sum_{i=1}^{m} Y_i$ is an unbiased estimator of $\mathcal{M}_\alpha(\boldsymbol{p})$, since

$$\mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}[Y_i] = \mathcal{M}_\alpha(\boldsymbol{p}).$$

Moreover,

$$\mathbf{Var}[Y_i] = \mathbf{E}\left[Y_i^2\right] - \mathbf{E}[Y_i]^2 = \sum_{j=1}^{n} \boldsymbol{p}[j]\,(\boldsymbol{p}[j])^{2\alpha-2} - \mathbf{E}[Y_i]^2 = \mathcal{M}_{2\alpha-1}(\boldsymbol{p}) - \mathcal{M}_\alpha^2(\boldsymbol{p}).$$

Since the Y_i's are also i.i.d. random variables,

$$\mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] = \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Var}\left[Y_i\right] = \frac{m}{m^2}\mathbf{Var}\left[Y\right] = \frac{1}{m}\left(\mathcal{M}_{2\alpha-1}\left(\boldsymbol{p}\right) - \mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right)\right).$$

We use the following fact from [7] to find an upper bound for the variance of our empirical estimator.

Fact 4.14. ([7], Lemma 1) *For $\alpha > 1$ and $0 \le \beta \le \alpha$*

$$\mathcal{M}_{\alpha+\beta}\left(\boldsymbol{p}\right) \le n^{(\alpha-1)(\alpha-\beta)/\alpha}\mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right) \quad .$$

By taking $\beta = \alpha - 1$, we get

$$\begin{aligned}\sigma^2 = \mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] &= \frac{1}{m}\left(\mathcal{M}_{2\alpha-1}\left(\boldsymbol{p}\right) - \mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right)\right)\\ &\le \frac{1}{m}\left(n^{(\alpha-1)/\alpha}\mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right) - \mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right)\right)\\ &= \frac{1}{m}\mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right)\left(n^{1-1/\alpha} - 1\right)\\ &\le \frac{\gamma^2}{100}\mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right).\end{aligned}$$

We obtain a similar upper bound for the standard deviation of our empirical estimator,

$$\sigma = \sqrt{\mathbf{Var}\left[\tfrac{1}{m}\sum_{i=1}^{m} Y_i\right]} \le \sqrt{\tfrac{\gamma^2}{100}\mathcal{M}_{\alpha}^{2}\left(\boldsymbol{p}\right)} \le \tfrac{\gamma}{10}\mathcal{M}_{\alpha}\left(\boldsymbol{p}\right).$$

By Chebyshev's inequality we have

$$\begin{aligned}&\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Y_i - \mathcal{M}_{\alpha}\left(\boldsymbol{p}\right)\right| > 10\sigma\right] \le \frac{1}{100} \Rightarrow\\ &\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Y_i - \mathcal{M}_{\alpha}\left(\boldsymbol{p}\right)\right| \le \gamma\mathcal{M}_{\alpha}\left(\boldsymbol{p}\right)\right] \ge 0.99\end{aligned}$$

Thus we can estimate $\mathcal{M}_\alpha(\boldsymbol{p})$ to a desired multiplicative accuracy with $O\left(\dfrac{n^{1-1/\alpha}}{\left(1-2^{(1-\alpha)\Delta}\right)^2}\right)$ queries, which ends the proof. □

```
Algorithm: ESTIMATOR VI
  • Fix $\gamma = 1 - 2^{(1-\alpha)\Delta}$ and $m = \left\lceil \frac{100n^{1-1/\alpha}}{\gamma^2} \right\rceil$
  • Draw $m$ independent samples $X_1, \dots, X_m$ from $\boldsymbol{p}$.
  • Compute $Y_i = (\boldsymbol{p}[X_i])^{\alpha-1}$ by evaluating PMF on $X_i$ for $1 \le i \le m$.
  • Output $\frac{1}{1-\alpha} \log \left( \frac{1}{m} \sum_{i=1}^{m} Y_i \right)$.
```

Figure 4.1. SAMP+PMF Estimator of Rényi Entropy (of degree $\alpha > 1$)

4.4. Support Size

Definition 4.15. *For a probability distribution* $\mathbf{p}$*, support size of* $\mathbf{p}$ *is defined as* $\operatorname{supp}(\boldsymbol{p}) = |\{i : \boldsymbol{p}[i] \neq 0\}|$*, the number of domain elements with nonzero probability.*

Recall that $H_0(\boldsymbol{p}) = \log \operatorname{supp}(\boldsymbol{p})$. The techniques described in Section 2.1 should be applicable to the distribution support size, since it is a symmetric property. In fact, Valiant and Valiant [19, 20] prove that for any positive constant $\epsilon < \frac{1}{4}$, estimating the support size of a distribution whose support members occur with probability at least $\frac{1}{n}$,[9] to within ±ϵn requires $\Theta(\frac{n}{\log n})$ independent samples. In addition, Canonne and Rubinfeld [10] show that $\Theta(1/\epsilon^2)$ SAMP+PMF queries are necessary (and sufficient) for estimating $\operatorname{supp}(\boldsymbol{p})$ with the guarantee that $\boldsymbol{p}[i] \geq \frac{1}{n}$ for all support members i, to within ±ϵn. We modify their proof to establish a matching lower bound for this task in the SAMP+CDF model.

Theorem 4.16. $\Omega\left(\dfrac{1}{\epsilon^2}\right)$ *SAMP+CDF queries are necessary to estimate (with high probability) the support size of an unknown distribution* $\mathbf{p}$ *on domain* $[n]$ *to within* ±ϵn*.*

[9] It is typical to assume that all elements in the support occur with probability at least $\frac{1}{n}$; since without such a lower bound it is impossible to estimate support size.

Proof. Assume that there exists a program $\mathcal{S}$ which can accomplish the task specified in the theorem statement with only $o\left(\frac{1}{\epsilon^2}\right)$ queries. Let us show how $\mathcal{S}$ can be used to determine whether a given a coin is fair, or comes up heads with probability $p_2 = \frac{1}{2} + \lambda$.

Set $\epsilon = \frac{\lambda}{6}$, and let n be the smallest even number satisfying $n \geq 10/\epsilon^2$. Partition the domain $[n]$ into $M = \frac{n}{2}$ blocks $I_1, \ldots, I_M$ where $I_m = \{2m-1, 2m\}$ for all $m \in [M]$. The construction of a probability distribution $\boldsymbol{p}$ based on coin flips is as follows:

- Throw the coin M times, with outcomes $X_1, \ldots, X_M$,
- Set $\boldsymbol{p}[2m-1] = \frac{2}{n}$ and $\boldsymbol{p}[2m] = 0$ if X_m is heads,
- Set $\boldsymbol{p}[2m-1] = \boldsymbol{p}[2m] = \frac{1}{n}$ if X_m is tails, for each $m \in [M]$.

Note that by construction $\boldsymbol{p}[2m-1] + \boldsymbol{p}[2m] = \frac{2}{n}$ for all $m \in [M]$. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the families of distributions constructed by the above process, using the fair and biased coin, respectively. Let $\boldsymbol{p_1}$ (respectively $\boldsymbol{p_2}$) be a probability distribution randomly chosen from $\mathcal{P}_1$ (respectively $\mathcal{P}_2$). Then

$$\begin{aligned}
\mathbf{E}\left[\operatorname{supp}\left(\boldsymbol{p_1}\right)\right] &= n - M\frac{1}{2} = n\left(1 - \frac{1/2}{2}\right) = \frac{3}{4}n, \\
\mathbf{E}\left[\operatorname{supp}\left(\boldsymbol{p_2}\right)\right] &= n - Mp_2 = n\left(1 - \frac{p_2}{2}\right) = n\left(\frac{3}{4} - \frac{\lambda}{2}\right) = \left(\frac{3}{4} - 3\epsilon\right)n,
\end{aligned}$$

and via the additive Chernoff bound,

$$\begin{aligned}
\Pr\left[\operatorname{supp}\left(\boldsymbol{p_1}\right) \leq \frac{3}{4}n - \frac{\epsilon}{2}n\right] &\leq e^{-\frac{\epsilon^2 n}{2}} \leq e^{-5} < \frac{1}{1000} \\
\Pr\left[\operatorname{supp}\left(\boldsymbol{p_2}\right) \geq \frac{3}{4}n - \frac{5\epsilon}{2}n\right] &\leq e^{-\frac{\epsilon^2 n}{2}} \leq e^{-5} < \frac{1}{1000}.
\end{aligned}$$

In other words, the resulting distributions will satisfy (with high probability) $\operatorname{supp}\left(\boldsymbol{p_1}\right) - \operatorname{supp}\left(\boldsymbol{p_2}\right) > 2\epsilon n$, distant enough for $\mathcal{S}$ to distinguish between two families.

As in our previous proofs, we could use $\mathcal{S}$ (if only it existed) to distinguish between the two possible coin types by using the coin for an on-the-fly construction of $\boldsymbol{p}$. As before, SAMP and CDF queries are answered by picking a block randomly, throwing the coin if the type of this block has not been fixed before, and returning the answer depending on the type of the block. Since $o\left(\frac{1}{\epsilon^2}\right)=o\left(\frac{1}{\lambda^2}\right)$ coin tosses would suffice for this task, we have reached a contradiction based on Lemma 4.2. □

4.5. Tsallis Entropy

Tsallis entropy [31], defined as

$$S_\alpha(\boldsymbol{p}) = \frac{\boldsymbol{k_B}}{\alpha - 1}\left(1 - \sum_{i=1}^{n} \left(\boldsymbol{p}\left[i\right]\right)^\alpha\right), \tag{4.8}$$

is a generalization of Boltzmann-Gibbs entropy where $\alpha \in \mathbb{R}$ and $\boldsymbol{k_B}$ is the Boltzmann constant. Harvey *et al.* [17] gives an algorithm to estimate the Tsallis entropy which is also used to approximate the Shannon entropy in the most general streaming model. Without loss of generality we focus on additively estimating the quantity $T_\alpha(\boldsymbol{p}) := \frac{S_\alpha(\boldsymbol{p})}{\boldsymbol{k_B}}$ which appears to be a generalization of Shannon entropy, easily derived via L'Hôpital's rule.

Lemma 4.17. *For any number $\alpha > 1$, there exists an algorithm estimating (with high probability) the Tsallis entropy of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with $O\left(\frac{1}{(\alpha-1)^2\Delta^2}\right)$ SAMP+PMF queries.*

Proof. Observe that for $\alpha > 1$

$$n^{1-\alpha} \leq \mathcal{M}_\alpha(\boldsymbol{p}) \leq 1 \quad \Rightarrow \quad 0 \leq T_\alpha\left(\boldsymbol{p}\right) \leq \frac{1}{\alpha - 1} - \frac{n^{1-\alpha}}{\alpha - 1}.$$

To estimate $T_\alpha\left(\boldsymbol{p}\right)$ to within additive error Δ, one needs to estimate $\mathcal{M}_\alpha(\boldsymbol{p})$ to within $\gamma = (\alpha - 1)\,\Delta$. Note that $\Delta < \frac{1}{\alpha-1}$, therefore, $\gamma < 1$ must satisfy for achieving nontrivial approximations of $T_\alpha\left(\boldsymbol{p}\right)$ and $\mathcal{M}_\alpha(\boldsymbol{p})$, respectively. We use the estimator constructed for Rényi entropy in Theorem 4.13.

Let $m = \lceil \frac{3}{(\alpha-1)^2\Delta^2} \rceil$ and draw m independent samples $X_1, \ldots, X_m$ from $\boldsymbol{p}$. Define $Y_i = (\boldsymbol{p}[X_i])^{\alpha-1}$ where $\boldsymbol{p}[X_i]$ is computed using a PMF query on X_i for $1 \le i \le m$. Recall that

$$\mathbf{E}[Y_i] = \sum_j (\boldsymbol{p}[j])^{\alpha-1} = \sum_j (\boldsymbol{p}[j])^{\alpha} = \mathcal{M}_\alpha(\boldsymbol{p}).$$

Obviously, $\frac{1}{m}\sum\limits_{i=1}^{m} Y_i$ is an unbiased estimator of $\mathcal{M}_\alpha(\boldsymbol{p})$. Observe that $Y_i \in [0,1]$ since $\alpha > 1$. By applying the Hoeffding bound, we get

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Y_i - \mathcal{M}_\alpha(\boldsymbol{p})\right| > \gamma\right] \le 2e^{-2\gamma^2 m}.$$

It easily follows that $m = O\left(\frac{1}{(\alpha-1)^2\Delta^2}\right) = O\left(\frac{1}{\gamma^2}\right)$ queries are sufficient to bound the right-hand side of the inequality. □

Jiao, Venkat and Weissman [32] construct an algorithm approximating Tsallis entropy $S_\alpha(\boldsymbol{p})$ of a distribution $\boldsymbol{p}$ to within additive error using $O\left(n^{2/\alpha-1}\right)$ SAMP queries for $1 < \alpha < 2$. More interestingly, they also prove that one requires only constant number of SAMP queries to additively estimate $S_\alpha(\boldsymbol{p})$ for $\alpha \ge 2$.[10]

[10] Acharya *et al.* [30] improve this result by constructing an algorithm for this task using only $O(1)$ samples for all $\alpha > 1$.

5. CONCLUSION

In this work, we investigate the task of additively estimating entropy in two settings based on two types of queries. A SAMP query takes no input and returns $x \in [n]$ with probability $\boldsymbol{p}[x]$; a PMF query takes as input $x \in [n]$ and returns the value $\boldsymbol{p}[x]$. The first setting is the SAMP model, where only SAMP queries are allowed. The second setting is the SAMP+PMF model in which both SAMP and PMF queries are utilized.

The motivation behind this work has both practical and theoretical reasons. Firstly, the SAMP+PMF model can be practical in many applications. For a concrete example, consider the Google n-gram database in which the frequency of each n-gram is published, and a random n-gram is easily obtained from the underlying text corpus. Secondly, the SAMP+PMF model is strongly related to the streaming model of computation which is an important field of computer science. Moreover, the SAMP+PMF model may illuminate the limitations of estimating entropy in the SAMP model.

We thoroughly analyzed the optimal bounds for estimating the Shannon entropy and the near-optimal bounds for estimating the Rényi entropy in the SAMP model. We described the exponentially faster algorithm constructed for estimating the Shannon entropy in the SAMP+PMF model. We established a matching lower bound for the estimation of the Shannon entropy $H\left(\boldsymbol{p}\right)$ in the SAMP+PMF model, $\Omega\left(\log^2 n\right)$. We gave optimal bounds for the estimation of the Rényi entropy $H_\alpha\left(\boldsymbol{p}\right)$ in the SAMP+PMF model, $\Theta\left(n^{1-1/\alpha}\right)$.

Apparently, PMF queries provided no advantage in estimating Rényi entropy for integer $\alpha > 1$, whereas they *were* advantageous for noninteger $\alpha > 1$. However, the proximity between $\Omega\left(n^{1-1/\alpha}\right)$ and $O\left(n\right)$, the lower and upper bound results for estimating Rényi entropy $H_\alpha\left(\boldsymbol{p}\right)$ for noninteger $\alpha > 1$ in the SAMP+PMF and SAMP models, respectively, indicated the amount of the advantage a PMF query added to the model. In addition, the exponential gap between $O\left(\log^2 n\right)$ and $\Omega\left(n^{1-1/\alpha}\right)$, the upper and lower bound results for estimating Shannon entropy and Rényi entropy in the SAMP+PMF model, respectively, implied the difficulty of the latter problem.

We proved that the bounds were easily extended to the SAMP+CDF model, where SAMP and CDF queries (given x, return $\sum_{y \leq x} \boldsymbol{p}[y]$) were allowed. We gave a matching lower bound for estimating support size to within $\pm\epsilon n$ in the SAMP+CDF model. Lastly, we constructed an algorithm for additively estimating Tsallis entropy $S_\alpha(\boldsymbol{p})$ using a constant number of SAMP+PMF queries.

One problem left open by our work is that of optimal bounds for estimating the Rényi entropy $H_\alpha(\boldsymbol{p})$ in the SAMP+PMF model for $\alpha < 1$. The work [7] shows that in the model where only SAMP are allowed, $\tilde{\tilde{\Omega}}\left(n^{1/\alpha}\right)$ queries are necessary when $0 < \alpha < 1$. It is interesting to ask whether there exists a sublinear algorithm for this task in the SAMP+PMF model.

Moreover, it is obvious that the SAMP+PMF model is superior to the SAMP model. However, degree of the superiority of the SAMP+PMF model is an open problem. In other words, how is the SAMP+PMF model affected if one has restricted number of such as $O(1)$ or $o(\log^2 n)$ PMF queries?

REFERENCES

1. Rubinfeld, R., “Taming Big Probability Distributions”, *XRDS: Crossroads, The ACM Magazine for Students*, Vol. 19, No. 1, pp. 24–28, 2012.

2. Canonne, C., *A Survey on Distribution Testing: Your Data is Big. But is It Blue?*, Tech. Rep. TR15-063, ECCC, 2015.

3. Paninski, L., “Estimation of Entropy and Mutual Information”, *Neural Computation*, Vol. 15, No. 6, pp. 1191–1253, 2003.

4. Paninski, L., “Estimating Entropy on m Bins Given Fewer than m Samples”, *IEEE Transactions on Information Theory*, Vol. 50, No. 9, pp. 2200–2203, 2004.

5. Valiant, P., “Testing Symmetric Properties of Distributions”, *SIAM Journal on Computing*, Vol. 40, No. 6, pp. 1927–1968, 2011.

6. Valiant, G. and P. Valiant, “Estimating the Unseen: An $n/\log(n)$-Sample Estimator for Entropy and Support Size, Shown Optimal via New CLTs”, *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pp. 685–694, 2011.

7. Acharya, J., A. Orlitsky, A. T. Suresh and H. Tyagi, “The Complexity of Estimating Rényi Entropy”, *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2015.

8. Kearns, M., Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire and L. Sellie, “On the Learnability of Discrete Distributions”, *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, pp. 273–282, 1994.

9. Batu, T., S. Dasgupta, R. Kumar and R. Rubinfeld, “The Complexity of Approximating the Entropy”, *SIAM Journal on Computing*, Vol. 35, No. 1, pp. 132–150, 2005.

10. Canonne, C. and R. Rubinfeld, *Testing Probability Distributions Underlying Aggregated Data*, Tech. Rep. TR14-021, Electronic Colloquium on Computational Complexity, 2014.

11. Alon, N., Y. Matias and M. Szegedy, “The Space Complexity of Approximating the Frequency Moments”, *Journal of Computer and System Sciences*, Vol. 58, No. 1, pp. 137–147, 1999.

12. Guha, S., A. McGregor and S. Venkatasubramanian, “Streaming and Sublinear Approximation of Entropy and Information Distances”, *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 733–742, ACM, 2006.

13. Lall, A., V. Sekar, M. Ogihara, J. Xu and H. Zhang, “Data Streaming Algorithms for Estimating Entropy of Network Traffic”, *Proceedings of ACM SIGMETRICS*, pp. 145–156, 2006.

14. Chakrabarti, A., K. Do Ba and S. Muthukrishnan, “Estimating Entropy and Entropy Norm on Data Streams”, *Internet Mathematics*, Vol. 3, No. 1, pp. 63–78, 2006.

15. Bhuvanagiri, L. and S. Ganguly, “Estimating Entropy over Data Streams”, *Proceedings of the 14th Annual European Symposium on Algorithms*, pp. 148–159, 2006.

16. Chakrabarti, A., G. Cormode and A. McGregor, “A Near-Optimal Algorithm for Computing the Entropy of a Stream”, pp. 328–335, 2007.

17. Harvey, N., J. Nelson and K. Onak, “Sketching and Streaming Entropy via Approximation Theory”, *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 489–498, 2008.

18. Caferov, C., B. Kaya, R. O’Donnell and A. C. C. Say, “Optimal Bounds for Estimating Entropy with PMF Queries”, *Proceedings of the 40th International Sym-*

posium on Mathematical Foundations of Computer Science, pp. 187–198, MFCS, 2015.

19. Valiant, G. and P. Valiant, *A CLT and Tight Lower Bounds for Estimating Entropy*, Tech. Rep. TR10-179, Electronic Colloquium on Computational Complexity, 2011.

20. Valiant, G. and P. Valiant, *Estimating the Unseen: A Sublinear-Sample Canonical Estimator of Distributions*, Tech. Rep. TR10-180, Electronic Colloquium on Computational Complexity, 2010.

21. Valiant, G. and P. Valiant, "The Power of Linear Estimators", *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, pp. 403–412, 2011.

22. Shannon, C. E., "A Mathematical Theory of Communication", *Bell System Technical Journal*, Vol. 27, No. 3, pp. 379—423, 1948.

23. Shenkin, P. S., B. Erman and L. D. Mastrandrea, "Information-Theoretical Entropy as a Measure of Sequence Variability", *Proteins*, Vol. 11, No. 4, pp. 297–313, 1991.

24. Valiant, P., *Testing Symmetric Properties of Distributions*, Tech. Rep. TR07-135, Electronic Colloquium on Computational Complexity, 2007.

25. Rényi, A., "On Measures of Entropy and Information", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume1: Contributions to the Theory of Statistics,*, pp. 547–561, 1961.

26. Oorschot, P. C. v. and M. J. Wiener, "Parallel Collision Search with Cryptanalytic Applications", *Journal of Cryptology*, Vol. 12, No. 1, pp. 1–28, 1999.

27. Batu, T., L. Fortnow, R. Rubinfeld, W. D. Smith and P. White, "Testing Closeness

of Discrete Distributions", *Journal of the ACM*, Vol. 60, No. 1, pp. 1927–1968, 2013.

28. Paninski, L., "A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data", *IEEE Transactions on Information Theory*, Vol. 54, No. 10, pp. 4750–4755, 2008.

29. Motahari, A. S., G. Bresler and D. N. Tse, "Information Theory of DNA Shotgun Sequencing", *IEEE Transactions on Information Theory*, Vol. 59, No. 10, pp. 6273–6289, 2013.

30. Acharya, J., A. Orlitsky, A. T. Suresh and H. Tyagi, *Estimating Rényi Entropy of Discrete Distributions*, Tech. Rep. arXiv:1408.1000v3, 2016.

31. Tsallis, C., *Possible Generalization of Boltzmann-Gibbs Statistics*, Tech. Rep. CBPF-NF-062/87, CBPF, 1987.

32. Jiao, J., K. Venkat and T. Weissman, "Maximum Likelihood Estimation of Functionals of Discrete Distributions", *CoRR*, Vol. abs/1406.6959, 2014.

APPENDIX A: INEQUALITIES

Theorem A.1. (Chebyshev's Inequality) *Let X be a real-valued random variable such that $\mathbf{Var}\left[X\right]$ is well-defined. Then, $\forall t > 0$,*

$$\Pr\left[\left|X - \mathbf{E}\left[X\right]\right| > t\right] \leq \frac{\mathbf{Var}\left[X\right]}{t^2}.$$

Theorem A.2. (Jensen's Inequality) *Let X be an integrable random variable and $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then,*

$$\mathbf{E}\left[\varphi\left(X\right)\right] \geq \varphi\left(\mathbf{E}\left[X\right]\right).$$

Theorem A.3. (Pinsker's Inequality) *Let $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ be two probability distributions on the universe U. Then*

$$\mathbf{KL}\left(\boldsymbol{p_1} \| \boldsymbol{p_2}\right) \geq \frac{1}{2\ln 2} \cdot \|\boldsymbol{p_1} - \boldsymbol{p_2}\|_1^2$$

Theorem A.4. (Exponential Inequality) *Let x be a real number. Then,*

$$1 + x \leq \left(1 + \frac{x}{n}\right)^n \leq e^x \quad \textit{for } n > 1, |x| \leq n.$$

Theorem A.5. (Chernoff Bound) *Let $X_1, \dots X_m$ be m independent random variables that take on values in $[0,1]$, where $\mathbf{E}\left[X_i\right] = p_i$, and $\sum\limits_{i=1}^{m} p_i = P$. For any $\gamma \in (0,1]$ we have*

$$\Pr\left[\sum_{i=1}^{m} X_i > \left(1+\gamma\right)P\right] \leq e^{-\gamma^2 P/3}, \;\; \Pr\left[\sum_{i=1}^{m} X_i < \left(1-\gamma\right)P\right] \leq e^{-\gamma^2 P/2}.$$

Theorem A.6. (Hoeffding Bound)[11] *Let $X_1, \ldots X_m$ be m independent random variables that take on values in $[0,1]$, where $\mathbf{E}[X_i] = p_i$, and $\sum\limits_{i=1}^{m} p_i = P$. For any $\gamma \in (0,1]$ we have*

$$\Pr\left[\sum_{i=1}^{m} X_i > P + \gamma m\right] \leq e^{-2\gamma^2 m}, \quad \Pr\left[\sum_{i=1}^{m} X_i < P - \gamma m\right] \leq e^{-2\gamma^2 m}.$$

[11] Usually, the Hoeffding bound is referred to as the additive Chernoff bound, whereas Theorem A.5 is referred to as the multiplicative Chernoff bound.