UNIVERSITY OF CALIFORNIA SAN DIEGO

Competitive and Universal Learning

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Yi Hao

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Sanjoy Dasgupta
Professor Massimo Franceschetti
Professor Tara Javidi
Professor Shachar Lovett

2021

The dissertation of Yi Hao is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents.

EPIGRAPH

Let there be light.

Genesis 1:3

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am grateful to my parents for their forever love, support, and kindness since I am simply nothing without them. I also understand that home is not where I live but where they miss me.

I want to thank my undergraduate thesis supervisor Raymond Yeung who showed me the doors to scientific research and helped me apply for graduate schools. I want to thank my previous colleague Ananda Theertha Suresh who helped me join the research group and find my dissertation topic. I want to thank my previous colleague Venkatadheeraj Pichapati with whom I had fruitful discussions and from whom I received encouraging words. I want to thank Ping Li for his supervision during my internship at Baidu Research. It was a terrific experience with two papers published at top machine learning conferences. I want to thank my current and other previous colleagues: Ayush Jain, Vaishakh Ravindrakumar, and Moein Falahatgar. We had lots of fun and excitement at the lab. And I will cherish the memories forever.

I have my highest respect and appreciation for my advisor Alon Orlitsky, the chair of my doctoral committee. It is always my great pleasure to be his student. Pretty much everything extraordinary I've heard about him has proven to be true. He is the brain, the heart, and the soul behind the yearly carnival - Information Theory and Applications (ITA) Workshop. If you've ever interacted with him, then you must know that he is also a man of quite large ambition, a genius of mentoring, and a master of his domain. He is an absolute perfectionist. As for him, there is nothing called 99.99% perfection. No surprise to find out is that he's also a constant paper-award winner, a man of integrity, and a Shannon Award laureate. He is a distinguished and brilliant researcher having profound visions for many fundamental questions. He is undoubtfully one of the most restless and dedicated scientists that I've ever known. Moreover, he is an uncommonly kind person, the enemy of none and a friend of all, the one and only Alon Orlitsky.

I am thankful to my committee members Sanjoy Dasgupta, Massimo Franceschetti, Tara Javidi, and Shachar Lovett. They are among the most incredible mentors and teachers that I know. I am fortunate to have completed 30 graduate courses from EE, CS, and MATH/STATS.

Chapter 2, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Yi Hao and Alon Orlitsky, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in *International Conference on Machine Learning*. Yi Hao and Alon Orlitsky, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in *International Conference on Machine Learning*. Yi Hao and Alon Orlitsky, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Yi Hao and Alon Orlitsky, 2020. The dissertation author was the primary investigator and author of this paper.

VITA

2015	B.Sc. in Mathematics, Chinese University of Hong Kong
2015	B.Eng. in Information Engineering, Chinese University of Hong Kong
2019	M.S. in Electrical Engineering (Communication Theory and Systems), University of California San Diego
2021	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California San Diego

PUBLICATIONS

Hao, Yi, and Alon Orlitsky. "Compressed Maximum Likelihood." *International Conference on Machine Learning (ICML)*, 2021. *Accepted*.

Hao, Yi, Ayush Jain, Alon Orlitsky, and Vaishakh Ravindrakumar. "SURF: A simple, universal, robust, fast distribution learning algorithm." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10881–10890. 2020.

Hao, Yi, and Ping Li. "Optimal prediction of the number of unseen species with multiplicity." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8553–8564. 2020.

Hao, Yi, and Alon Orlitsky. "Profile entropy: A fundamental measure for the learnability and compressibility of distributions." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6947–6958. 2020.

Hao, Yi, and Alon Orlitsky. "Data amplification: Instance-optimal property estimation." *International Conference on Machine Learning (ICML)*, pp. 4049–4059. PMLR, 2020.

Hao, Yi, and Ping Li. "Bessel smoothing and multi-distribution property estimation." *Conference on Learning Theory (COLT)*, pp. 1817–1876. PMLR, 2020.

Hao, Yi, and Alon Orlitsky. "The broad optimality of profile maximum likelihood." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10989–11001. 2019.

Hao, Yi, and Alon Orlitsky. "Unified sample-optimal property estimation in near-linear time." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11104–11114. 2019.

Hao, Yi, and Alon Orlitsky. "Doubly competitive distribution estimation" *International Conference on Machine Learning (ICML)*, pp. 2614–2623. PMLR, 2019.

Hao, Yi, Alon Orlitsky, and Venkatadheeraj Pichapati. "On learning Markov chains." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 646–655. 2018.

Hao, Yi, Alon Orlitsky, Ananda T. Suresh, and Yihong Wu. "Data amplification: A unified and competitive approach to property estimation." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8848–8857. 2018.

Hao, Yi, and Alon Orlitsky. "Adaptive estimation of generalized distance to uniformity." *IEEE International Symposium on Information Theory (ISIT)*, pp. 1076–1080. IEEE, 2018.

Falahatgar, Moein, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. "Maxing and ranking with few assumptions." *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7063–7073. 2017.

Zhao, Jinyu*, Yi Hao*, and Cyrus Rashtchian. "Unsupervised Embedding of Hierarchical Structure in Euclidean Space." *arXiv preprint arXiv:2010.16055*, 2020. *In submission*.

ABSTRACT OF THE DISSERTATION

Competitive and Universal Learning

by

Yi Hao

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2021

Professor Alon Orlitsky, Chair

Modern data science calls for statistical inference algorithms that are both data-efficient and computation-efficient. We design and analyze methods that 1) outperform existing algorithms in the high-dimensional setting; 2) are as simple as possible but efficiently computable.

One line of our work aims to bridge different fields and offer unified solutions to multiple questions. Consider the following canonical statistical inference tasks: distribution estimation, functional estimation, and property testing, sharing the model that provides sample access to an unknown discrete distribution. In a recent paper in NeurIPS '19, we showed that a single, simple, and unified estimator – profile maximum likelihood (PML), and its near-linear time computable variant are sample-optimal for estimating multiple distribution attributes. The result covers 1)

any appropriately Lipschitz and additively separable functionals; 2) sorted distribution; 3) Rényi entropy; 4) ℓ_2 distance to the uniform distribution, yielding an optimal tester for distributions' closeness. This work makes PML the first unified sample- and time-optimal method for the learning tasks mentioned above. A single algorithm with such broad applicability is *universal*.

Another line of our work focuses on instance-optimal learning that designs algorithms with near-optimal guarantees for every possible data input. A flagship problem is distribution estimation over discrete or continuous domains, where ordering and geometry play an essential role. Going beyond worst-case guarantees, researchers designed algorithms that compete with a genie estimator that knows the actual distribution but is reasonably restricted. To obtain state-of-the-art algorithms for both tasks, we leveraged the simple but nontrivial idea of "data binning". For discrete settings, we group symbols that appear the same number of times. And for continuous settings, we partition the real domain and separate symbols according to pre-designed local quantiles. The respective algorithms run in near-linear-time, achieve the best-known estimation guarantees regarding the genie estimators, and appear in ICML'19 and NeurIPS '20. A genie-like algorithm adaptive to almost every data sample is *competitive*.

We present a comprehensive understanding of universal and competitive algorithms for multiple fundamental learning problems. Our ideas and techniques may shed light on key challenges in modern data science and numerous applications beyond the scope of this dissertation.

Chapter 1

Introduction

The primary purpose of our research lies in designing and analyzing statistical inference algorithms, with a focus on several challenges emerging from modern data science: dimensionality and structure, sample and time efficiency, and algorithmic adaptiveness to data complexity.

To meet these challenges and narrow the gap between theory and practice, we strive to propose and study methods that: 1) outperform existing algorithms in the high-dimensional setting or data-sparse regime; 2) are as simple as possible but efficiently computable; 3) possess near-optimal guarantees for every possible data input; 4) bridge and unify approaches in different fields. Below, we illustrate these points through concrete results and examples and introduce some general notation for subsequent sections along the way.

1.1 Universal Learning

Consider the following canonical statistical inference problems: distribution estimation, functional estimation, and distribution property testing. They all share the setting that provides access to samples from an unknown discrete distribution p. Respectively, they aim to learn p or its probability multiset, approximate a (symmetric) functional $f(p)$, and test whether p equals or is far from a known distribution. Motivated by the diverse applications ranging from

ecology to language modeling, researchers in learning, information theory, computer science have studied these fundamental statistical questions in the high-dimensional regime during the past two decades, and proposed numerous methods.

Wouldn't it be nice to have a unified and straightforward approach that optimally addresses all three problems? We asked this question after working on some specific sub-questions such as entropy estimation. We first observed that these inference tasks share the same sufficient statistic – sample profile, defined as the multiset of empirical counts of different symbols. Concurrently, a paper by our group showed that for a given sample profile, the respective (profile) maximum likelihood distribution (PML) serves as the optimal plug-in estimators for four specific functionals, including Shannon entropy and support size. Mathematically, given a sample $X^n \sim p$, the multiplicity of a symbol x is the number of times x appearing in X^n. The sample profile, $\varphi(X^n)$, is the multiset of multiplicities of different symbols in X^n. And the respective PML estimate is the distribution that maximizes the probability of observing a length-n sample with profile $\varphi(X^n)$.

Impressed by how natural and intrinsic these arguments are, we realized that the PML method holds the potential of being the first universal, sample- and time-optimal approach in high dimensions for all the learning tasks mentioned above.

With the conjecture in mind, we showed that the simple PML (plug-in) estimator and its near-linear time computable variant are sample-optimal for estimating 1) any appropriately Lipschitz and additively separable functionals, such as Shannon entropy and ℓ_1 distance to the uniform; 2) sorted probability distribution under ℓ_1 distance; 3) Rényi entropy of arbitrary order $\alpha > 3/4$; 4) ℓ_2 distance to the uniform distribution, yielding an optimal tester for distributions' closeness. Moreover, one can compute the PML once and simultaneously perform all these tasks while maintaining high confidence in the estimation and testing accuracy.

The respective research paper [HO19a], titled "the broad optimality of profile maximum likelihood", appeared in NeurIPS 2019 as a spotlight presentation. At the heart of the work is a broad class of explicit estimators with optimal sample efficiency and exponentially small error

probabilities, accompanied by a set of duality and chaining arguments.

1.2 Instance-Optimal Learning

The success of PML reveals the power of symmetry and maximum likelihood, but not every learning problem has such elegant properties. A flagship problem is distribution estimation, either over discrete or continuous domains, where ordering and geometry play an essential role. Classical formulations aim to design min-max estimators. And it is well-known that the naive empirical distribution and its add-constant variants are optimal under several performance measures, including ℓ_1 distance and KL divergence. However, practical distributions are rarely the worst possible and often quite simple, such as power-law and exponential, rendering the min-max approach overly pessimistic, and its estimators, typically sub-optimal in practice.

Researchers turned to competitive or instance-optimal formulations in the past decade, seeking algorithms that compete with a genie estimator that knows the actual distribution but is reasonably restricted. For discrete settings, the genie estimator is assumed to be natural. Namely, it assigns the same probability to symbols appearing the equal number of times. For continuous distributions, the genie estimator is chosen from an expressive class such as piecewise polynomials. The former formulation reflects the best human-designed estimator without any prior knowledge on the distribution, while the latter intends to exploit distribution structure assumptions at hand.

Given these formulations, one designs learning algorithms whose estimation error, for every distribution, is at most a constant multiple of that of the best genie, plus a distribution-free slack term that vanishes with the sample size. As the genie incurs an error that represents the limit of estimation for each distribution instance, research works along this line have been focused on determining the best constant factor and the lowest error slack term. Mathematically, for any p, let $\mathrm{OPT}_n(p)$ be the n-sample loss of the optimal genie. Then, one wants to find an estimator $\hat{p}$

whose n-sample estimation loss $\ell_n(\hat{p}, p)$ is at most $c \cdot \mathrm{OPT}_n(p) + f(n)$ for every p, where c is a constant, and $f(n)$ is an error function that vanishes with n, typically at an $O(1/\sqrt{n})$ rate.

To obtain state-of-the-art algorithms for both problems, we leveraged the simple but nontrivial idea of "data binning" that puts observed symbols into different categories. Intuitively, a single observation's behavior is often dominated by randomness, but that of a group of symbols would be robust and predictable.

For discrete settings, we group symbols appearing an equal number of times. Our method estimates the total probability mass in each group through an improved Good-Turing smoothing scheme, then distributes the mass equally among the group's symbols. For continuous settings, we compete with a standard and powerful genie that approximates the underlying density with the best degree-d piecewise polynomial(s). Our method greedily partitions the real domain into disjoint intervals, then groups symbols within each according to pre-designed local quantiles.

Through the simple yet powerful binning idea, we created algorithms that achieve the best-known competitive estimation guarantees. Recall from the last section that the profile of X^n is the multiset $\varphi(X^n)$ of the multiplicities of observed symbols. Define the dimension of a profile φ as the number $\mathcal{D}(\varphi)$ of distinct multiplicities. Over discrete domains and under KL divergence, our approach achieves an optimal factor of $c = 1$ and learns each distribution to the optimal slack $f(n)$ of order $n^{-1} \cdot \mathcal{D}(\varphi(X^n)) \leq 2n^{-1/2}$, with high probability. Over the real domain and under ℓ_1 distance, our method attains the best constants c among all existing estimators that are efficiently computable for multiple d values. For example, both the log-concave and monotone density mixtures are well approximated by piecewise linear ($d = 1$) curves, for which our method achieves a factor of $c = 2.25$ without knowing the number of mixture components, while the best current algorithm yields $c \geq 9$.

For discrete distribution learning, the method is near-linear time computable, appearing in an ICML '19 paper [HO19b]. We established the optimality of the respective slack error term $n^{-1} \cdot \mathcal{D}(\varphi(X^n))$ in a subsequent work [HO20b] in NeurIPS '20. For continuous density

estimation, our algorithm runs in polynomial time for any d, presented in a separate NeurIPS '20 paper [HJOR20].

1.3 Data Amplification

A paradox frequently encountered in designing new learning algorithms is that we want to approach an optimal algorithm's performance, but have few clues for how the optimal learner behaves, as it seems that otherwise, the problem is resolved. To get around this obstacle, we proposed an intuitive and useful methodology, data amplification, with the idea of amplifying the amount of data at hand for simple algorithms to boost their performance.

Consider the following generic learning formulation: given a sample X^n from an unknown distribution p, learn some of its attributes $\mathcal{T}(p)$. For distribution estimation, the attribute $\mathcal{T}(p)$ is the distribution itself; for real functional estimation, $\mathcal{T}(p)$ is a functional value $f(p)$; for parameter estimation, $\mathcal{T}(p)$ is a collection of parameters controlling the behavior of the distribution. Also, let $\mathcal{A}$ be any algorithm that takes X^n to approximate $\mathcal{T}(p)$ by $\mathcal{A}(X^n)$.

The method of data amplification takes as a benchmark a natural and commonly used estimator $\mathcal{A}$, which should be relatively easy to analyze. For a large class of learning tasks, it constructs an n-sample estimator that instance-by-instance performs as well as $\mathcal{A}$ would perform with a much larger sample size $\mathcal{A}(X^{na})$, where $a > 1$ is a properly chosen factor. It therefore effectively "amplifies" the amount of data available compared to common estimation techniques.

During the past three years, we applied the data amplification method to a sequence of fundamental learning problems and obtained the best results for most of them. A cornerstone example is Shannon entropy estimation. The best-known and most commonly used estimator is the empirical (plug-in) estimator that evaluates the empirical distribution's entropy. For distributions over a large, size-k alphabet, the expected worst-case error achieved by the n-sample empirical estimator is of order k/n, whenever $n \geq k$. In the past decade, a sequence of research works

showed that for $n \geq k/\log k$, more sophisticated estimators achieve the min-max performance of order $k/(n\log n)$, demonstrating a logarithmic improvement.

Hence, in terms of the worst-case performance, the n-sample min-max entropy estimator is as good as the $n\log n$-sample empirical estimator. Applying the data amplification method with $\mathcal{A}$ being the empirical estimator and $a = \log n$, we showed that this n-to-$n\log n$ phenomenon holds for each and every discrete distribution, regardless of the potentially unknown alphabet size. As a direct corollary, we established that in the $n \ll k/\log k$ data-sparse regime, the min-max estimation rate becomes order $\log(1 + k/(n\log n))$, revealing a surprising elbow effect.

Following the method's general applicability, the n-to-$n\log n$ data amplification phenomenon holds for a broad class of functionals beyond Shannon entropy. For example, the above result covers any functional in the $\sum_x g_x(p_x)$ form, where p_x is the probability of symbol x, and g_x is a given 1-Lipschitz real function over $[0,1]$. The proposed algorithms require only a sample, run in near-linear time, and amplify the effective sample size by order of magnitude for every distribution, with high probability.

Inspired by federated learning with decentralized data from different local sources, we also considered a more general data amplification setting where one has sample access to multiple distributions and wishes to understand their relationships. Numerous essential learning problems fall into this category and can be formulated as tasks of estimating multi-distribution functionals. For example, mixture testing asks whether one distribution is (close to) a linear combination of the rest, and independence testing queries if one distribution is near the others' product.

Perhaps surprisingly, nearly all existing works address cases with at most two distributions, and little is known about how to design sample- and time-efficient algorithms for such high-dimensional functionals. Fortunately, data amplification again shed light on the new direction.

For a given sample size and distribution, the empirical-distribution plug-in estimator possesses a simple form and smooth expectation (viewed as a function of probabilities). Leveraging this fact, we applied a smoothing method based on Bessel functions and derived a family of

estimators that match the empirical estimator's performance with access to more samples. For several essential statistical learning tasks, including the mixture and independence testing, we obtained the first efficiently computable estimator whose sample complexity is sub-linear in the distributions' alphabet sizes.

Recently, we applied data amplification to a generalization of the well-known unseen species problem, initially studied by Fisher in the early 1940s. The new method led to estimators with optimal sample and time complexities, and recovered many notable results in prior works. The respective paper [HL20b] is accepted to NeurIPS '20 as a spotlight presentation.

The generic approach for learning entropy, support size, and many other single-distribution functionals appears in NeurIPS '18 [HOSW18] and ICML '20 [HO20a], while paper [HO20a] presents a stronger guarantee. The linear-time algorithm for learning the relationship among multiple data sources appears in COLT '20 [HL20a].

1.4 Thesis Organization

The rest of this thesis is organized as follows:

- Chapter 2 establishes the universal optimality of the *profile maximum likelihood (PML)* estimator for numerous fundamental learning tasks, ranging from estimating functionals to approximating sorted distributions to testing distribution closeness.

- Chapter 3 designs a competitive distribution estimator that outperforms the widely renowned *Good-Turing smoothing technique*, both theoretically and experimentally.

- Chapter 4 derives an instance-optimal estimator for entropy and many other Lipschitz functionals. For every possible distribution, the method *amplifies the effective data size* and uses only $n/\log n$ observations to approach the performance of the empirical plug-in estimator corresponding to n observations.

- Chapter 5 proposes *profile entropy*, a novel complexity measure unifying the concepts of estimation, inference, and compression for sequential models. The results built on this measure connect and complete those in preceding chapters.

Chapter 2

The Broad Optimality of Profile Maximum Likelihood

2.1 Introduction

A distribution p over a discrete alphabet $\mathcal{X}$ of size k corresponds to an element of the standard simplex

$$\Delta_{\mathcal{X}} := \left\{ p \in \mathbb{R}^k_{\geq 0} : \sum_{x \in \mathcal{X}} p(x) = 1 \right\}.$$

A distribution *property* is a mapping $f : \Delta_{\mathcal{X}} \to \mathbb{R}$ associating a real value with each distribution. A distribution property f is *symmetric* if it is invariant under domain-symbol permutations. A symmetric property is *additive*, i.e., additively separable, if it can be written as $f(p) := \sum_x f(p(x))$, where for simplicity we use f to denote both the property and the corresponding real function.

Many important symmetric properties are additive. For example,

- **Support size** $S(p) := \sum_x \mathbb{1}_{p(x)>0}$, a fundamental quantity arising in the study of vocabulary size [ET76, McN73, TE87], population estimation [Goo53, ML07], and database studies [HNSS95].

- **Support coverage** $C_m(p) := \sum_x (1 - (1 - p(x))^m)$, where m is a given parameter, the expected number of distinct elements observed in a sample of size m, arising in biological [Cha84, KLR99] and ecological [Cha84, CC14, CL92, CCG$^+$12] research;

- **Shannon entropy** $H(p) := -\sum_x p(x) \log p(x)$, the primary measure of information [CT12, Sha48] with numerous applications to machine learning [Bre15, CL68, QKC13] and neuroscience [GK02, MS95];

- **Distance to uniformity** $D(p) := \|p - p_u\|_1$, where p_u is the uniform distribution over $\Delta_{\mathcal{X}}$, a property being central to the field of distribution property testing [BFR$^+$00, BFF$^+$01, Can17, Ron10].

Besides being additive and symmetric, these four properties have yet another attribute in common. Under the appropriate interpretation, they are also all 1-Lipschitz. Specifically, for two distributions $p, q \in \Delta_{\mathcal{X}}$, let $\Gamma_{p,q}$ be the collection of distributions over $\mathcal{X} \times \mathcal{X}$ with marginals p and q on the first and second factors respectively. The *relative earth-mover distance* [VV11b], between p and q is

$$R(p,q) := \inf_{\gamma \in \Gamma_{p,q}} \mathop{\mathbb{E}}_{(X,Y)\sim\gamma} \left| \log \frac{p(X)}{q(Y)} \right|.$$

One can verify [VV11b, VV16] that H, D, and $\tilde{C}_m := C_m/m$ are all 1-Lipschitz on the metric space $(\Delta_{\mathcal{X}}, R)$, and $\tilde{S} := S/k$ is 1-Lipschitz over $(\Delta_{\geq 1/k}, R)$, the set of distributions in $\Delta_{\mathcal{X}}$ whose nonzero probabilities are at least $1/k$. We will study all such Lipschitz properties in later sections.

An important symmetric non-additive property is *Rényi entropy*, a well-known measure of randomness with numerous applications to unsupervised learning [JHE$^+$03, Xu99] and image registration [MHGM00, NHZC06]. For a distribution $p \in \Delta_{\mathcal{X}}$ and a non-negative real parameter $\alpha \neq 1$, the α-*Rényi entropy* [Rén61] of p is $H_\alpha(p) := (1-\alpha)^{-1} \log (\sum_x p_x^\alpha)$. In particular, denoted by $H_1(p) := \lim_{\alpha \to 1} H_\alpha(p)$, the 1-*Rényi entropy* is exactly Shannon entropy [Rén61].

2.1.1 Problems of Interest

In this work, we consider three fundamental statistical learning problems concerning the estimation and testing of distributions and their properties.

(Sorted) Distribution Estimation

A natural learning problem is to estimate an unknown distribution $p \in \Delta_{\mathcal{X}}$ from an i.i.d. sample $X^n \sim p$. For any two distributions $p, q \in \Delta_{\mathcal{X}}$, let $\ell(p,q)$ be the *loss* when we approximate p by q. A *distribution estimator* $\hat{p} : \mathcal{X}^* \to \Delta_{\mathcal{X}}$ associates every sequence $x^n \in \mathcal{X}^*$ with a distribution $\hat{p}(x^n)$. We measure the performance of an estimator by its *sample complexity*

$$n(\hat{p}, \varepsilon, \delta) := \min\{n : \forall p \in \Delta_{\mathcal{X}}, \Pr_{X^n \sim p}(\ell(p, \hat{p}(X^n)) \geq \varepsilon) \leq \delta\},$$

the smallest sample size that $\hat{p}$ requires to estimate all distributions in $\Delta_{\mathcal{X}}$ to a desired accuracy $\varepsilon > 0$, with error probability $\delta \in (0,1)$. The sample complexity of distribution estimation is

$$n(\varepsilon, \delta) := \min\{n(\hat{p}, \varepsilon, \delta) : \hat{p} : \mathcal{X}^* \to \Delta_{\mathcal{X}}\},$$

the lowest sample complexity of any estimator. For simplicity, we will omit δ when $\delta = 1/3$.

For a distribution $p \in \Delta_{\mathcal{X}}$, we denote by $\{p\}$ the multiset of its probabilities. The *sorted* ℓ_1 *distance* between two distributions $p, q \in \Delta_{\mathcal{X}}$ is

$$\ell_1^{\downarrow}(p,q) := \min_{p' \in \Delta_{\mathcal{X}} : \{p'\} = \{p\}} \left\| p' - q \right\|_1,$$

the smallest ℓ_1 distance between q and any sorted version of p. As illustrated in Section 2.D.1, this is essentially the 1-Wasserstein distance between uniform measures on the probability multisets $\{p\}$ and $\{q\}$. We consider both the sorted and unsorted ℓ_1 distances.

Property Estimation

Often we would like to estimate a given property f of an unknown distribution $p \in \Delta_{\mathcal{X}}$ based on a sample $X^n \sim p$. A *property estimator* is a mapping $\hat{f} : \mathcal{X}^* \to \mathbb{R}$. Analogously, the *sample complexity* of $\hat{f}$ in estimating f over a set $\mathcal{P} \subset \Delta_{\mathcal{X}}$ is

$$n_f(\hat{f}, \mathcal{P}, \varepsilon, \delta) := \min\{n : \forall p \in \mathcal{P}, \Pr_{X^n \sim p}(|\hat{f}(X^n) - f(p)| \geq \varepsilon) \leq \delta\},$$

the smallest sample size that $\hat{f}$ requires to estimate f with accuracy ε and confidence $1-\delta$, for all distributions in $\mathcal{P}$. The sample complexity of estimating f over $\mathcal{P}$ is

$$n_f(\mathcal{P}, \varepsilon, \delta) := \min\{n_f(\hat{f}, \mathcal{P}, \varepsilon, \delta) : \hat{f} : \mathcal{X}^* \to \mathbb{R}\},$$

the lowest sample complexity of any estimator. For simplicity, we will omit $\mathcal{P}$ when $\mathcal{P} = \Delta_{\mathcal{X}}$, and omit δ when $\delta = 1/3$. The standard "median trick" shows that $\log(1/\delta) \cdot n_f(\mathcal{P}, \varepsilon) \geq \Omega(n_f(\mathcal{P}, \varepsilon, \delta))$. By convention, an estimator $\hat{f}$ is *sample-optimal* if $n_f(\hat{f}, \mathcal{P}, \varepsilon) = \Theta(n_f(\mathcal{P}, \varepsilon))$.

Property Testing: Identity Testing

A closely related problem is distribution property testing, of which identity testing is the most fundamental and well-studied [Can17, Gol17]. Given an error parameter ε, a distribution q, and a sample X^n from an unknown distribution p, *identity testing* aims to distinguish between the null hypothesis

$$H_0 : p = q$$

and the alternative hypothesis

$$H_1 : \|p - q\|_1 \geq \varepsilon.$$

A *property tester* is a mapping $\hat{t} : \mathcal{X}^* \to \{0, 1\}$, indicating whether H_0 or H_1 is accepted.

Analogous to the two formulations above, the *sample complexity* of $\hat{t}$ is

$$n_q(\hat{t},\varepsilon,\delta) := \min\{n : \forall i \in \{0,1\} \text{ and } \forall p \in H_i, \Pr_{X^n\sim p}(\hat{t}(X^n) \neq i) \leq \delta\},$$

and the sample complexity of identity testing with respect to q is

$$n_q(\varepsilon,\delta) := \min\{n(\hat{t},\varepsilon,\delta) : \hat{t} : \mathcal{X}^* \to \{0,1\}\}.$$

Again, when $\delta = 1/3$, we omit δ. For $q = p_u$, the problem is also known as *uniformity testing*.

2.1.2 Profile Maximum Likelihood

The *multiplicity* of a symbol $x \in \mathcal{X}$ in a sequence $x^n := x_1,\ldots,x_n \in \mathcal{X}^*$ is $\mu_x(x^n) := |\{j : x_j = x, 1 \leq j \leq n\}|$, the number of times x appears in x^n. These multiplicities induce an *empirical distribution* $p_\mu(x^n)$ that associates a probability $\mu_x(x^n)/n$ with each symbol $x \in \mathcal{X}$.

The *prevalence* of an integer $i \geq 0$ in x^n is the number $\varphi_i(x^n)$ of symbols appearing i times in x^n. For known $\mathcal{X}$, the value of φ_0 can be deduced from the remaining multiplicities, hence we define the *profile* of x^n to be $\varphi(x^n) = (\varphi_1(x^n),\ldots,\varphi_n(x^n))$, the vector of all positive prevalences. For example, $\varphi(\textit{alfalfa}) = (0,2,1,0,0,0,0)$. Note that the profile of x^n also corresponds to the multiset of multiplicities of distinct symbols in x^n.

For a distribution $p \in \Delta_\mathcal{X}$, let

$$p(x^n) := \Pr_{X^n\sim p}(X^n = x^n)$$

be the probability of observing a sequence x^n under i.i.d. sampling from p, and let

$$p(\varphi) := \sum_{y^n:\varphi(y^n)=\varphi} p(y^n)$$

be the probability of observing a profile φ. While the sequence maximum likelihood estimator maps a sequence to its empirical distribution, which maximizes the sequence probability $p(x^n)$, the *profile maximum likelihood (PML)* estimator [OSVZ04] over a set $\mathcal{P} \subseteq \Delta_{\mathcal{X}}$ maps each profile φ to a distribution

$$p_\varphi := \arg\max_p p(\varphi)$$

that maximizes the profile probability. Relaxing the optimization objective, for any $\beta \in (0,1)$, a β-*approximate PML* estimator [ADOS17] maps each profile φ to a distribution p^β_φ such that $p^\beta_\varphi(\varphi) \geq \beta \cdot p_\varphi(\varphi)$.

Originating from the principle of maximum likelihood, PML was proved [ADJ$^+$12a, ADOS17, ABKS18, AGZ17, Das12, OSVZ04] to possess a number of useful attributes, such as existence over finite discrete domains, majorization by empirical distributions, consistency for distribution estimation under both sorted and unsorted ℓ_1 distances, and competitiveness to other profile-based estimators.

Let ε be an error parameter and f *be one of the four properties* in the introduction. Set $n := n_f(\varepsilon)$. Recently, [ADOS17] showed that for some absolute constant $c' > 0$, if $c < c'$ and $\varepsilon \geq n^{-c}$, then a plug-in estimator for f, using an $\exp(-n^{1-\Theta(c)})$-approximate PML, is sample-optimal. Motivated by this result, [CSS19b] constructed an explicit $\exp(-O(n^{2/3}\log^3 n))$-approximate PML (APML) whose computation time is near-linear in n. Combined, these results provide a unified, sample-optimal, and near-linear-time computable plug-in estimator for the four properties.

2.2 New Results and Implications

2.2.1 New Results

Additive Property Estimation

Recall that for any property f, the expression $n_f(\varepsilon)$ denotes the smallest sample size required by any estimator to achieve accuracy ε with confidence $2/3$, for all distributions in $\Delta_{\mathcal{X}}$. Let f be an additive symmetric property that is 1-Lipschitz on $(\Delta_{\mathcal{X}}, R)$. Let $\varepsilon > 0$ and $n \geq n_f(\varepsilon)$ be error and sampling parameters. For an absolute constant $c \in (10^{-2}, 10^{-1})$, if $\varepsilon \geq n^{-c}$,

Theorem 1. *The PML plug-in estimator, when given a sample of size* $4n$ *from any distribution* $p \in \Delta_{\mathcal{X}}$*, will estimate* $f(p)$ *up to an error of* $(2+o(1))\varepsilon$*, with probability at least* $1-\exp(-4\sqrt{n})$.

For a different $c > 0$, Theorem 1 also holds for APML, which is always near-linear-time computable [CSS19b].

Rényi Entropy Estimation

For $\mathcal{X}$ of size k and any $p \in \Delta_{\mathcal{X}}$, it is well-known that $H_\alpha(p) \in [0, \log k]$. The following theorems characterize the performance of the PML plug-in estimator in estimating Rényi entropy.

For any distribution $p \in \Delta_{\mathcal{X}}$, error $\varepsilon \in (0,1)$, absolute constant $\lambda \in (0, 0.1)$, and sampling parameter n, draw a sample $X^n \sim p$ and denote its profile by φ. Then for sufficiently large k,

Theorem 2. *For* $\alpha \in (3/4, 1)$*, if* $n = \Omega_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha}\log k))$,

$$\Pr\left(|H_\alpha(p_\varphi) - H_\alpha(p)| \geq \varepsilon\right) \leq \exp(-\sqrt{n}).$$

Theorem 3. *For non-integer* $\alpha > 1$*, if* $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$,

$$\Pr\left(|H_\alpha(p_\varphi) - H_\alpha(p)| \geq \varepsilon\right) \leq \exp(-n^{1-\lambda}).$$

Theorem 4. *For integer* $\alpha > 1$*, if* $n = \Omega_\alpha(k^{1-1/\alpha}(\varepsilon^2 \log(1/\varepsilon))^{-(1+\alpha)})$ *and* $H_\alpha(p) \le (\log n)/4$*,*

$$\Pr(|H_\alpha(p_\varphi) - H_\alpha(p)| \ge \varepsilon) \le 1/3.$$

Replacing $3/4$ by $5/6$, Theorem 2 also holds for APML with a better probability bound $\exp(-n^{2/3})$. In addition, Theorem 3 holds for APML without any modifications.

Sorted Distribution Estimation

Let c be the absolute constant defined just before Theorem 1. For any distribution $p \in \Delta_\mathcal{X}$, error $\varepsilon \in (0,1)$, and sampling parameter n, draw a sample $X^n \sim p$ and denote its profile by φ.

Theorem 5. *If* $n = \Omega(n(\varepsilon)) = \Omega\left(k/(\varepsilon^2 \log k)\right)$ *and* $\varepsilon \ge n^{-c}$*,*

$$\Pr(\ell_1^\uparrow(p_\varphi, p) \ge \varepsilon) \le \exp(-\Omega(\sqrt{n})).$$

For a *different* $c > 0$, Theorem 5 also holds for APML with a significantly better probability bound $\exp(-n^{2/3})$.

Identity Testing

The recent works of [DK16] and [Gol16] provided a procedure reducing identity testing to uniformity testing, while modifying the desired accuracy and alphabet size by only absolute constant factors. Hence below we consider uniformity testing.

The uniformity tester T_{PML} shown in Figure 2.1 is purely based on PML and satisfies

Theorem 6. *If* $\varepsilon = \tilde{\Omega}(k^{-1/4})$ *and* $n = \tilde{\Omega}(\sqrt{k}/\varepsilon^2)$*, the tester* $T_{PML}(X^n)$ *will be correct with probability at least* $1 - k^{-2}$*. The tester also distinguishes between* $p = p_u$ *and* $\|p - p_u\|_2 \ge \varepsilon/\sqrt{k}$*.*

The $\tilde{\Omega}(\cdot)$ notation only hides logarithmic factors of k. The tester T_{PML} is near-optimal as for uniform distribution p_u, [DGPP18] yields an $\Omega(\sqrt{k \log k}/\varepsilon^2)$ lower bound on $n_{p_u}(\varepsilon, k^{-2})$.

For space considerations, we postpone proofs and additional results to the Appendices. The rest of the paper is organized as follows. Section 2.2.2 presents several immediate implications of the above theorems. Section 2.3 and Section 2.4 illustrate PML's theoretical and practical advantages by comparing it to existing methods for a variety of learning tasks. Section 2.5 concludes the paper and outlines multiple promising future directions.

Input: parameters $k, \varepsilon,$ and a sample $X^n \sim p$ with profile φ.
if $\max_x \mu_x(X^n) \geq 3\max\{1, n/k\}\log k$ **then return** 1;
elif $\left\|p_\varphi - p_u\right\|_2 \geq 3\varepsilon/(4\sqrt{k})$ **then return** 1;
else return 0.

Figure 2.1: Uniformity tester T_{PML}

2.2.2 Implications

Several immediate implications are in order. We say that a plug-in estimator is *universally sample-optimal* for estimating symmetric properties if there exist absolute positive constants c_1, c_2 and c_3, such that for any 1-Lipschitz property on $(\Delta_\mathcal{X}, R)$, with probability $\geq 9/10$, the plug-in estimator uses just c_1 times the sample size n required by the minimax estimator to achieve c_2 times its error, whenever this error is at least n^{-c_3}.

Note that the "1-Lipschitz property" class can be replaced by other general property classes, but not by those containing only a few specific properties, since "universal" means "applicable to all cases".

Theorem 1 makes PML the *first* plug-in estimator that is *universally sample-optimal* for a broad class of distribution properties. In particular, Theorem 1 also covers the four properties considered in [ADOS17]. To see this, as mentioned in the introduction, $\tilde{C}_m$, H, and D are 1-Lipschitz on $(\Delta_\mathcal{X}, R)$; as for $\tilde{S}$, the following result [ADOS17] relates it to $\tilde{C}_m$ for distributions in $\Delta_{\geq 1/k}$, and proves PML's optimality.

Lemma 1. *For any* $\varepsilon > 0$, $m = k\log(1/\varepsilon)$, *and* $p \in \Delta_{\geq 1/k}$,

$$|\tilde{S}(p) - \tilde{C}_m(p)\log(1/\varepsilon)| \leq \varepsilon.$$

The theorem also applies to many other properties. As an example [VV11b], given an integer $s > 0$, let $f_s(x) := \min\{x, |x - 1/s|\}$. Then to within a factor of two, $f_s(p) := \sum_x f_s(p_x)$ approximates the ℓ_1 distance between any distribution p and the closest uniform distribution in $\Delta_{\mathcal{X}}$ of support size s.

In Section 2.3.1 we compare Theorem 1 with existing results and present more of its strong implications.

Theorem 2 and 3 imply that for all non-integer $\alpha > 3/4$ (resp. $\alpha > 5/6$), the PML (resp. APML) plug-in estimator achieves a sample complexity better than the best currently known [AOST16]. This makes both the PML and APML plug-in estimators the state-of-the-art algorithms for estimating non-integer order Rényi entropy. See Section 2.3.2 for a review of known results, and see Section 2.3.2 for a detailed comparison between existing methods and ours.

Theorem 4 shows that for all integer $\alpha > 1$, the sample complexity of the PML plug-in estimator has optimal $k^{1-1/\alpha}$ dependence [AOST16, OS17] on the alphabet size.

Theorem 5 makes APML the first distribution estimator under sorted ℓ_1 distance that is both near-linear-time computable and sample-optimal for a range of desired accuracy ε beyond inverse polylogarithmic of n. In comparison, existing algorithms [ADJ$^+$12a, HJW18, VV11a] either run in polynomial time in the sample sizes, or are only known to achieve optimal sample complexity for $\varepsilon = \Omega(1/\sqrt{\log n})$, which is essentially different from the applicable range of $\varepsilon \geq n^{-\Theta(1)}$ in Theorem 5. We provide a more detailed comparison in Section 2.3.3.

Theorem 6 provides the first PML-based uniformity tester with near-optimal sample complexity. As stated, the tester also distinguishes between $p = p_u$ and $\|p - p_u\|_2 \geq \varepsilon/\sqrt{k}$. This is a stronger guarantee since by the Cauchy-Schwarz inequality, $\|p - p_u\|_1 \geq \varepsilon$ implies

$\|p - p_u\|_2 \geq \varepsilon/\sqrt{k}$.

Note that several other uniformity testers in the literature (see Section 2.3.4) also provide the same ℓ_2 testing guarantee, since all of them are essentially counting sample collisions, i.e., the number of location pairs such that the sample points at those locations are equal.

2.3 Related Work and Comparisons

2.3.1 Additive Property Estimation

The study of additive property estimation dates back at least half a century [Car69, Goo53, GT56] and has steadily grown over the years. For any additive symmetric property f and sequence x^n, the simplest and most widely-used approach uses the *empirical (plug-in)* estimator $\hat{f}^E(x^n) := f(p_\mu(x^n))$ that evaluates f at the empirical distribution. While the empirical estimator performs well in the large-sample regime, modern data science applications often concern high-dimensional data, for which more involved methods have yielded property estimators that are more sample-efficient. For example, for relatively large k and for f being $\tilde{S}$, $\tilde{C}_m$, H, or D, recent research [JVHW15, OSW16, VV11a, VV11b, WY16, WY19] showed that the empirical estimator is optimal up to logarithmic factors, namely $n_f(\mathcal{P}, \varepsilon) = \Theta_\varepsilon(n_f(\hat{f}^E, \mathcal{P}, \varepsilon)/\log n_f(\hat{f}^E, \mathcal{P}, \varepsilon))$, where $\mathcal{P}$ is $\Delta_{\geq 1/k}$ for $\tilde{S}$, and is $\Delta_\mathcal{X}$ for the other properties.

Below we classify the methods for deriving the corresponding sample-optimal estimators into two categories: plug-in and approximation, and provide a high-level description. For simplicity of illustration, we assume that $\varepsilon \in (0, 1]$.

The *plug-in* approach essentially estimates the unknown distribution multiset, which suffices for computing any symmetric properties. Besides the empirical and PML estimators, [ET76] proposed a linear-programming approach that finds a multiset estimate consistent with the sample's profile. This approach was then adapted and analyzed by [VV11a, VV11a], yielding plug-in estimators that achieve near-optimal sample complexities for H and $\tilde{S}$, and optimal sample

complexity for D, when ε is relatively large.

The *approximation* approach modifies non-smooth segments of the probability function to correct the bias of empirical estimators. A popular modification is to replace those non-smooth segments by their low-degree polynomial approximations and then estimate the modified function. For several properties including the above four and *power sum* $P_\alpha(p) := \sum_x p_x^\alpha$, where α is a given parameter, this approach yields property-dependent estimators [JVHW15, OSW16, WY16, WY19] that are sample-optimal for all ε.

More recently, [ADOS17] proved the aforementioned results on PML estimator and made it the first unified, sample-optimal plug-in estimator for $\tilde{S}$, $\tilde{C}_m$, H and D and relatively large ε. Following these advances, [HJW18] refined the linear-programming approach and designed a plug-in estimator that implicitly performs polynomial approximation and is sample-optimal for H, $\tilde{S}$, and P_α with $\alpha < 1$, when ε is relatively large.

Comparison I: Theorem 1 and Related Property-Estimation Work

In terms of the estimator's theoretical guarantee, Theorem 1 is essentially the same as [VV11b]. However, for each property, k, and n, [VV11b] solves a different linear program and constructs a new estimator, which takes polynomial time. On the other hand, both the PML estimator and its near-linear-time computable variant, once computed, can be used to accurately estimate exponentially many properties that are 1-Lipschitz on $(\Delta_\mathcal{X}, R)$. A similar comparison holds between the PML method and the approximation approach, while the latter is provably sample-optimal for only a few properties. In addition, Theorem 1 shows that the PML estimator often achieves the optimal sample complexity up to a small constant factor, which is a desired estimator attribute shared by some, but not all approximation-based estimators [JVHW15, OSW16, WY16, WY19].

In term of the method and proof technique, Theorem 1 is closest to [ADOS17]. On the other hand, [ADOS17] establishes the optimality of PML for only four properties, while

our result covers a much broader property class. In addition, both the above mentioned "small constant factor" attribute, and the confidence boost from $2/3$ to $1-\exp(-4\sqrt{n})$ are unique contributions of this work. The PML plug-in approach is also close in flavor to the plug-in estimators in [VV11a, VV11a] and their refinement in [HJW18]. On the other hand, as pointed out previously, these plug-in estimators are provably sample-optimal for only a few properties. More specifically, for estimating H, $\tilde{S}$, and $\tilde{C}_m$, the plug-in estimators in [VV11a, VV11a] achieve sub-optimal sample complexities with regard to the desired accuracy ε; and the estimation guarantee in [HJW18] is provided in terms of the approximation errors of $\tilde{O}(\sqrt{n})$ polynomials that are not directly related to the optimal sample complexities.

2.3.2 Rényi Entropy Estimation

Motivated by the wide applications of Rényi entropy, heuristic estimators were proposed and studied in the physics literature following [Gra88], and asymptotically consistent estimators were presented and analyzed in the statistical learning literature [KLS12, XE10]. For the special case of 1-Rényi (or Shannon) entropy, the works of [VV11a, VV11b] determined the sample complexity to be $n_f(\varepsilon) = \Theta(k/(\varepsilon \log k))$.

For general α-Rényi entropy, the best-known results in [AOST16] state that for integer and non-integer α values, the corresponding sample complexities $n_f(\varepsilon, \delta)$ are $O_\alpha(k^{1-1/\alpha}\log(1/\delta)/\varepsilon^2)$ and $O_\alpha(k^{\min\{1/\alpha,1\}}\log(1/\delta)/(\varepsilon^{1/\alpha}\log k))$, respectively. The upper bounds for integer α are achieved by an estimator that corrects the bias of the empirical plug-in estimator. To achieve the upper bounds for non-integer α values, one needs to compute some best polynomial approximation of z^α, whose degree and domain both depend on n, and construct a more involved estimator using the approximation approach [JVHW15, WY16] mentioned in Section 2.3.1.

Comparison II: Theorem 2 to 4 and Related Rényi-Entropy-Estimation Work

Our result shows that a single PML estimate suffices to estimate the Rényi entropy of different orders α. Such adaptiveness to the order parameter is a significant advantage of PML over existing methods. For example, by Theorem 3 and the union bound, one can use a *single* APML or PML to accurately approximate exponentially many non-integer order Rényi entropy values, yet still maintains an overall confidence of $1-\exp(-k^{0.9})$. By comparison, the estimation heuristic in [AOST16] requires different polynomial-based estimators for different α values. In particular, to construct each estimator, one needs to compute some best polynomial approximation of z^α, which is not known to admit a closed-form formula for $\alpha \notin \mathbb{Z}$. Furthermore, even for a single α and with a sample size $\sqrt{k}$ times larger, such estimator is not known to achieve the same level of confidence as PML or APML.

As for the theoretical guarantees, the sample-complexity upper bounds in both Theorem 2 and 3 are better than those mentioned in the previous section. More specifically, for any $\alpha \in (3/4, 1)$ and $\delta \geq \exp(-k^{-0.5})$, Theorem 2 shows that $n_f(\varepsilon, \delta) = O_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha}\log k))$. Analogously, for any non-integer $\alpha > 1$ and $\delta \geq \exp(-k^{-0.9})$, Theorem 3 shows that $n_f(\varepsilon, \delta) = O_\alpha(k/(\varepsilon^{1/\alpha}\log k))$. Both bounds are better than the best currently known by a $\log(1/\delta)$ factor.

2.3.3 (Sorted) Distribution Estimation

Estimating large-alphabet distributions from their samples is a fundamental statistical learning tenet. Over the past few decades, distribution estimation has found numerous applications, ranging from natural language modeling [CG99] to biological research [AIS+08], and has been studied extensively. Under the classical ℓ_1 and KL losses, existing research [BS04, KOPS15] showed that the corresponding sample complexities $n(\varepsilon)$ are $\Theta(k/\varepsilon^2)$ and $\Theta(k/\varepsilon)$, respectively. Several recent works have investigated the analogous formulation under sorted ℓ_1 distance, and revealed a lower sample complexity of $n(\varepsilon) = \Theta(k/(\varepsilon^2 \log k))$. Specifically, under certain conditions, [VV11a, HJW18] derived sample-optimal estimators using linear programming, and

[ADJ$^+$12a, Das12] showed that PML achieves a sub-optimal $O(k/(\varepsilon^{2.1}\log k))$ sample complexity for relatively large ε.

Comparison III: Theorem 5 and Related Distribution-Estimation Work

We compare our results with existing ones from three different perspectives.

Applicable parameter ranges: As shown by [HJW18], for $\varepsilon \ll n^{-1/3}$, the simple empirical estimator is already sample-optimal. Hence we consider the parammeter range $\varepsilon = \Omega(n^{-1/3})$. For the results in [ADJ$^+$12a, Das12] and [VV11a] to hold, we would need ε to be at least $\Omega(1/\sqrt{\log n})$. On the other hand, Theorem 5 shows that PML and APML are sample-optimal for ε larger than $n^{-\Theta(1)}$. Here, the gap is exponentially large. The result in [HJW18] applies to the whole range $\varepsilon = \Omega(n^{-1/3})$, which is larger than the applicable range of our results.

Time complexity: Both the APML and the estimator in [VV11a] are near-linear-time computable in the sample sizes, while the estimator in [HJW18] would require polynomial time to be computed.

Statistical confidence: The PML and APML achieve the desired accuracy with an error probability at most $\exp(-\Omega(\sqrt{n}))$. On the contrary, the estimator in [HJW18] is known to achieve an error probability that decreases only as $O(n^{-3})$. The gap is again exponentially large. The estimator in [VV11a] admits a better error probability bound of $\exp(-n^{0.02})$, which is still far from ours.

2.3.4 Identity Testing

Initiated by the work of [GR00], identity testing is arguably one of the most important and widely-studied problem in distribution property testing. Over the past two decades, a sequence of works [ADK15, BFF$^+$01, DK16, DKN15, DGPP18, GR00, Pan08, VV17] have addressed the sample complexity of this problem and proposed testers with a variety of guarantees. In

particular, applying a coincidence-based tester, [Pan08] determined the sample complexity of uniformity testing up to constant factors; utilizing a variant of the Pearson chi-squared statistic, [VV17] resolved the general identity testing problem. For an overview of related results, we refer interested readers to [Can17] and [Gol17]. The contribution of this work is mainly showing that PML, is a unified sample-optimal approach for several related problems, and as shown in Theorem 6, also provides a near-optimal tester for this important testing problem.

2.4 Experiments and Distribution Estimation

A number of different approaches have been taken to computing the PML and its approximations. Among the existing works, [ADM$^+$10] considered exact algebraic computation, [OSS$^+$04, OSVZ04] designed an EM algorithm with MCMC acceleration, [Von12, Von14] proposed a Bethe approximation heuristic, [AGZ17] introduced a sieved PML estimator and a stochastic approximation of the associated EM algorithm, and [PJW17] derived a dynamic programming approach. Notably and recently, for a sample size n, [CSS19b] constructed an explicit $\exp(-O(n^{2/3}\log^3 n))$-approximate PML whose computation time is near-linear in n.

In Section 2.A we introduce a variant of the MCMC-EM algorithm in [Pan12] and demonstrate the exceptional efficacy of PML on a variety of learning tasks through experiments. In particular, we derive a new distribution estimator for (unsorted) ℓ_1 distance by combining the proposed PML computation algorithm with the denoising procedure in [VV16] and a novel missing mass estimator. As shown below, the proposed distribution estimator has the state-of-the-art performance.

In Figures 2.2, samples are generated according to six distributions of the same support size $k = 5{,}000$. Details about these distributions can be found in Section 2.A.2. The sample size n (horizontal axis) ranges from 10,000 to 100,000, and the vertical axis reflects the (unsorted) ℓ_1 distance between the true distribution and the estimates, averaged over 30 independent tri-

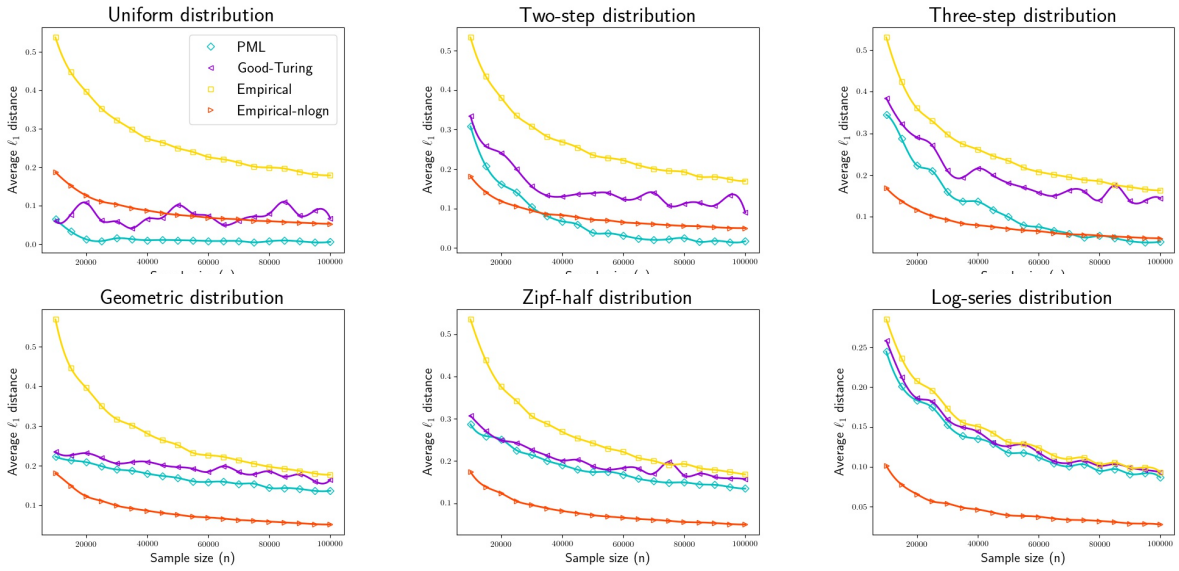


Figure 2.2: Distribution estimation under ℓ_1 distance

als. We compare our estimator with three different ones: the improved Good-Turing estimator in [OS15, HO19b], which is provably instance-by-instance near-optimal [OS15], the empirical estimator, serving as a baseline, and the empirical estimator with a larger $n \log n$ sample size. Note that $\log n$ is roughly 11. As shown in [OS15], the improved Good-Turing estimator substantially outperformed other estimators such as the Laplace (add-1) estimator, the Braess-Sauer estimator [BS04], and the Krichevsky-Trofimov estimator [KT81]. Hence we do not include those estimators here. The following plots showed that our proposed estimator further outperformed the improved Good-Turing estimator in all the experiments.

2.5 Conclusion and Future Directions

We studied three fundamental problems in statistical learning: distribution estimation, property estimation, and property testing. We established the profile maximum likelihood (PML) as the first universally sample-optimal approach for several important learning tasks: distribution estimation under the sorted ℓ_1 distance, additive property estimation, Rényi entropy estimation,

and identity testing. Several future directions are promising. We believe that neither the factor of 4 in the sample size in Theorem 1, nor the lower bounds on ε in Theorem 1, 5, and 6 are necessary. In other words, the PML approach is universally sample-optimal for these tasks in all ranges of parameters. It is also of interest to extend the PML's optimality to estimating symmetric properties not covered by Theorem 1 to 4, such as *generalized distance to uniformity* [BC17, HOSW18], the ℓ_1 distance between the unknown distribution and the closest uniform distribution over an arbitrary subset of $\mathcal{X}$.

Another important direction is *competitive (or instance-optimal) property estimation*. It should be noted that all the referenced works including this paper are of the worst-case nature, namely, designing estimators with near-optimal worst-case performances. On the contrary, practical and natural distributions often possess simple structures, and are rarely the worst possible. To address this discrepancy, the recent work [HO20a, HOSW18] took a competitive approach and constructed estimators whose performances are adaptive to the simplicity of the underlying distributions. Specifically, for any property in a broad class and *every* distribution in $\Delta_{\mathcal{X}}$, the expected error of the proposed estimator with a sample of size $n/\log n$ is at most that of the empirical estimator with a sample of size n, pluses a distribution-free vanishing function of n. These results not only cover $\tilde{S}$, $\tilde{C}_m$, H, and D, for which the $\log n$-factor is optimal up to constants, but also apply to any *non-symmetric additive* property $\sum_x f_x(p_x)$ where f_x is 1-Lipschitz for all $x \in \mathcal{X}$, such as the ℓ_1-distance to a given distribution. It would be of interest to study the optimality of the PML approach under this formulation as well. Readers interested in estimating non-symmetric properties may also find the paper [HO19c] helpful.

2.6 Acknowledgments

Chapter 2, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Yi Hao and Alon Orlitsky, 2019. The dissertation author was the

primary investigator and author of this paper.

2.A Numerical Experiments

A number of different approaches have been taken to computing the PML and its approximations. Among the existing works, [ADM+10] considered exact algebraic computation, [OSS+04, OSVZ04] designed an EM algorithm with MCMC acceleration, [Von12, Von14] proposed a Bethe approximation heuristic, [AGZ17] introduced a sieved PML estimator and a stochastic approximation of the associated EM algorithm, and [PJW17] derived a dynamic programming approach. Notably and recently, for a sample size n, [CSS19b] constructed an explicit $\exp(-O(n^{2/3}\log^3 n))$-approximate PML whose computation time is near-linear in n.

In this section, we first introduce a variant of the MCMC-EM algorithm in [OSS+04, OSVZ04, Pan12] and then demonstrate the efficacy of PML on a variety of learning tasks through experiments.

2.A.1 MCMC-EM Algorithm Variant

To approximate PML, the work [OSS+04] proposed an MCMC-EM algorithm, where MCMC and EM stand for Markov chain Monte Carlo and expectation maximization, respectively. A sketch of the original MCMC-EM algorithm can be found in [OSS+04], and a detailed description is available in Chapter 6 of [Pan12]. The EM part uses a simple iteration procedure to update the distribution estimates. One can show [Pan12] that it is equivalent to the conventional *generalized gradient ascent method*. The MCMC part exploits local properties of the update process and accelerates the EM computation. Below we present a variant of this algorithm that often runs faster and is more accurate.

Step 1: We separate the large and small multiplicities. Define a threshold parameter $\tau := 1.5\log^2 n$ and suppress X^n in $p_\mu(X^n)$ for simplicity. For symbols x with $\mu_x(X^n) \geq \tau$, estimate

their probabilities by $p_\mu(x) = \mu_x(X^n)/n$ and remove them from the sample. Denote the collection of removed symbols by R and the remaining sample sequence by X^r. In the subsequent steps, we apply the EM-MCMC algorithm to X^r.

The idea is simple: By the Chernoff-type bound for binomial random variables, with high probability, the empirical frequency $\mu_x(X^n)/n$ of a large-multiplicity symbol x is very close to its mean value $p(x)$. Hence for large-multiplicity symbols we can simply use the empirical estimates and focus on estimating the probabilities of small-multiplicity symbols. This is similar to initializing the EM algorithm by the empirical distribution and fixing the large probability estimates through the iterations. However, the approach described here is more efficient.

Step 2: We determine a proper alphabet size for the output distribution of the EM algorithm. If the true value k is provided, then we simply use $k - |R|$. Otherwise, we apply the following support size estimator [ADOS17] to X^r:

$$\hat{S}(X^r) := \sum_{j \geq 1} (1 - (-(t-1))^j \Pr(L \geq j)) \cdot \varphi_j(X^r),$$

where $t = \log r$ and L is an independent binomial random variable with support size $\lceil \frac{1}{2} \log_2(\frac{rt^2}{t-1}) \rceil$ and success probability $(t+1)^{-1}$. For any ε larger than an absolute constant, estimator $\hat{S}$ achieves the optimal sample complexity $n_f(\Delta_{\geq 1/k}, \varepsilon)$ in estimating support size, up to constant factors [ADOS17].

Step 3: Apply the MCMC-EM algorithm in [OSS$^+$04, Pan12] to $\varphi(X^r)$ with the output alphabet size determined in the previous step, and denote the resulting distribution estimate by p_r. (In the experiments, we perform the EM iteration for 30 times.) Intuitively, this estimate corresponds to the conditional distribution given that the next observation is a symbol with small probability.

Step 4: Let $T_\mu := \sum_{x \in R} p_\mu(x)$ be the total probability of the large-multiplicity symbols. Treat p_r as a vector and let $p'_r := (1 - T_r) \cdot p_r$. For every symbol $x \in R$, append $p_\mu(x)$ to p'_r, and

return the resulting vector. Note that this vector corresponds to a valid discrete distribution.

Algorithm Code

The implementation of our algorithm is available at `https://github.com/ucsdyi/PML`.

For computational efficiency, the program code for the original MCMC-EM algorithm in [OSS+04, Pan12] is written in C++, with a file name "MCMCEM.cpp". The program code for other functions is written in Python3. Note that to execute the program, one should have a 64-bit Windows/Linux system with Python3 installed (64-bit version). In addition, we also use functions provided by "NumPy" and "SciPy", while the latter is not crucial and can be removed by modifying the code slightly.

Our implementation also makes use of "ctypes", a *built-in* foreign language library for Python that allows us to call C++ functions directly. Note that before calling C++ functions in Python, we need to compile the corresponding C++ source files into DLLs or shared libraries. We have compiled and included two such files, one is "MCMCEM.so", the other is "MCMCEM.dll".

Functions in "MCMCEM.cpp" can be used separately. To compute a PML estimate, simply call the function "int PML(int MAXSZ=10000, int maximum_EM=20, int EM_n=100)", where the first parameter specifies an upper bound on the support size of the output distribution, the second provides the maximum number of EM iteration, and the last corresponds to the sample size n. This function takes as input a local file called "proFile", which contains the profile vector $\varphi(X^n)$ in the format of "1 4 7 10 ...". Specifically, the file "proFile" consists of only space-separated non-negative integers, and the i-th integer represents the value of $\varphi_i(X^n)$. The output is a vector of length at most MAXSZ, and is stored in another local file called "PMLFile". Each line of "PMLFile" contains a non-negative number, corresponding to a probability estimate.

To perform experiments and save the plots to the directory containing the code, simply execute the file "Main.py". To avoid further complication, the code compares our estimator with only three other estimators: empirical, empirical with a larger $n \log n$ sample size, and improved

Good-Turing [OS15] (for distribution estimation under unsorted ℓ_1 distance). The implementation covers all the distributions described in the next section. One can test any of these distributions by including it in "D_List" of the "main()" function. The implementation also covers a variety of learning tasks, such as distribution estimation under sorted and unsorted ℓ_1 distances, and property estimation for Shannon entropy, α-Rényi entropy, support coverage, and support size.

Finally, functions related to distribution and sample generation are available in file "Samples.py". Others including the property computation functions, the sorted and unsorted ℓ_1 distance functions, and the previously-described support size estimator, are in file "Functions.py".

2.A.2 Experiment Distributions

In the following experiments, samples are generated according to six distributions with the same support size $k = 5{,}000$.

Three of them have finite support by definition: uniform distribution, two-step distribution with half the symbols having probability $2/(5k)$ and the other half have probability $8/(5k)$, and a three-step distribution with one third the symbols having probability $3/(13k)$, another third having probability $9/(13k)$, and the remaining having probability $27/(13k)$.

The other three distributions are over $\{i \in \mathbb{Z} : i \geq 1\}$, and are truncated at $i = 5{,}000$ and re-normalized: geometric distribution with parameter $g = 1/k$ satisfying $p_i \propto (1-g)^i$, Zipf distribution with parameter $1/2$ satisfying $p_i \propto i^{-1/2}$, and log-series distribution with parameter $\gamma = 2/k$ satisfying $p_i \propto (1-\gamma)^i/i$.

2.A.3 Experiment Results and Details

The proposed PML approximation algorithm has exceptional performance.

Distribution Estimation under ℓ_1 Distance

We derive a new distribution estimator under the (unsorted) ℓ_1 distance by combining the proposed PML computation algorithm with the denoising procedure in [VV16] and a missing mass estimator [OS15].

First we describe this distribution estimator, which takes a sample X^n from some unknown distribution p. An optional input is $\mathcal{X}$, the underlying alphabet.

Step 1: Apply the PML computation algorithm described in Section 2.A.1 to X^n, and denote the returned vector, consisting of non-negative real numbers that sum to 1, by V.

Step 2: Employ the following variant of the denoising procedure in [VV16]. Arbitrarily remove a total probability mass of $\log^{-2} n$ from entries of the vector V without making any entry negative. Then for each $j \leq \log^2 n$, augment the vector by $n/(j\log^4 n)$ entries of probability j/n. For every multiplicity $\mu \geq 1$ appearing in the sample, assign to all symbols appearing μ times the following probability value. If $\mu \geq \log^2 n$, simply assign to each of these symbols the empirical estimate μ/n; otherwise, temporally associate a weight of $\mathrm{bin}(n,v,\mu) := \binom{n}{\mu}(1-v)^{n-\mu}v^{\mu}$ with each entry v in V, and assign to each of these symbols the current weighted median of V.

Step 3: If $\mathcal{X}$ is available, we can estimate the total probability mass $M(X^n) := \sum_{x\in\mathcal{X}} \mathbb{1}_{x\notin X^n}$ of the unseen symbols (a.k.a., the *missing mass*) by the following estimator:

$$\hat{M}(X^n) := \frac{\varphi_1(X^n)}{\sum_j (j\varphi_j(X^n)\mathbb{1}_{j>\varphi_{j+1}} + (j+1)\varphi_{j+1}(X^n)\mathbb{1}_{j\leq\varphi_{j+1}})}.$$

We equally distribute this probability mass estimate among symbols that do not appear in the sample. As shown below, this distribution estimator achieves the state-of-the-art performance.

In Figures 2.3, the horizontal axis reflects the sample size n, ranging from 10,000 to 100,000, and the vertical axis reflects the (unsorted) ℓ_1 distance between the true distribution and the estimates, averaged over 30 independent trials. We compare our estimator with three others: the improved Good-Turing estimator [OS15, HO19b], the empirical estimator, serving as

a baseline, and the empirical estimator with a larger $n \log n$ sample size. Note that $\log n$ is roughly 11. As shown in [OS15], the improved Good-Turing estimator is provably instance-by-instance near-optimal and substantially outperforms other estimators such as the Laplace (add-1) estimator, the Braess-Sauer estimator [BS04], and the Krichevsky-Trofimov estimator [KT81]. Hence we do not include those estimators below.

As the following plots show, our proposed estimator outperformed the improved Good-Turing estimator in all experiments.

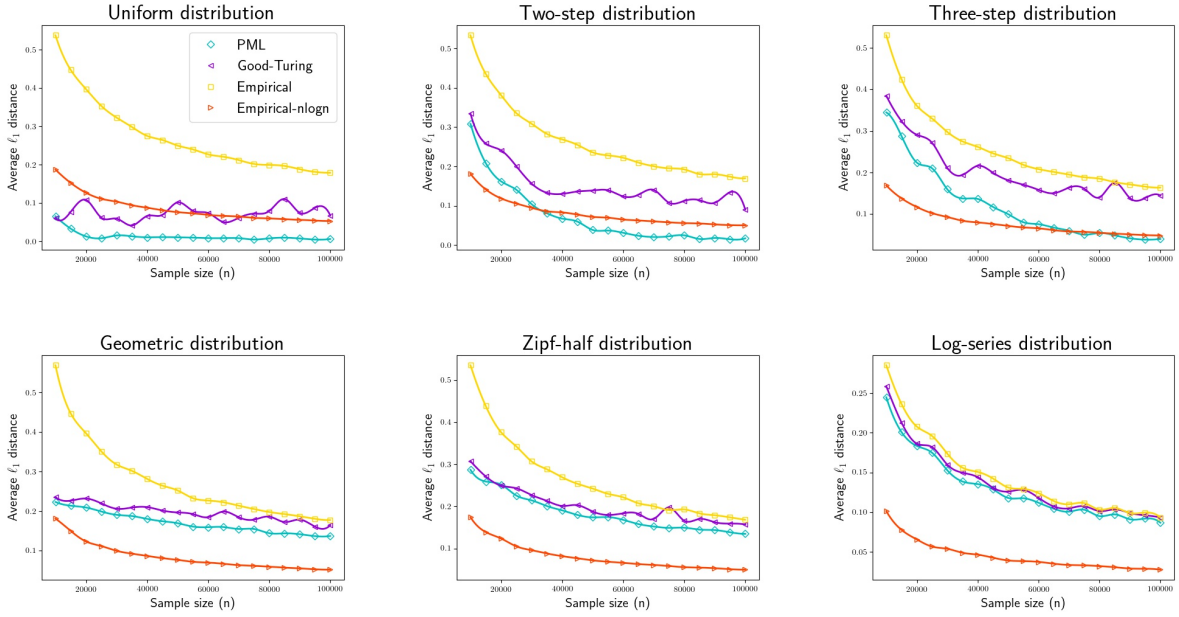


Figure 2.3: Distribution estimation under ℓ_1 distance

Distribution Estimation under Sorted ℓ_1 Distance

In Figure 2.4, the sample size n ranges from 2,000 to 20,000, and the vertical axis reflects the sorted ℓ_1 distance between the true distribution and the estimates, averaged over 30 independent trials. We compare our estimator with that proposed by [VV11a] that utilizes linear programming, with the empirical estimator, and with the empirical estimator with a larger $n \log n$ sample size.

We do not include the estimator in [HJW18] since there is no implementation available, and as pointed out by the recent work of [VKVK19] (page 7), the approach in [HJW18] "is quite unwieldy. It involves significant parameter tuning and special treatment for the edge cases." and "Some techniques . . . are quite crude and likely lose large constant factors both in theory and in practice."

As shown in Figure 2.4, with the exception of uniform distribution, where the estimator in [VV11a] (VV-LP) is the best and PML is the closest second, the PML estimator outperforms VV-LP for all other tested distributions. As the underlying distribution becomes more skewed, the improvement of PML over VV-LP grows. For the log-series distribution, the performance of VV-LP is even worse than the empirical estimator. Additionally, the plots also demonstrate that PML has a more stable performance than VV-LP.

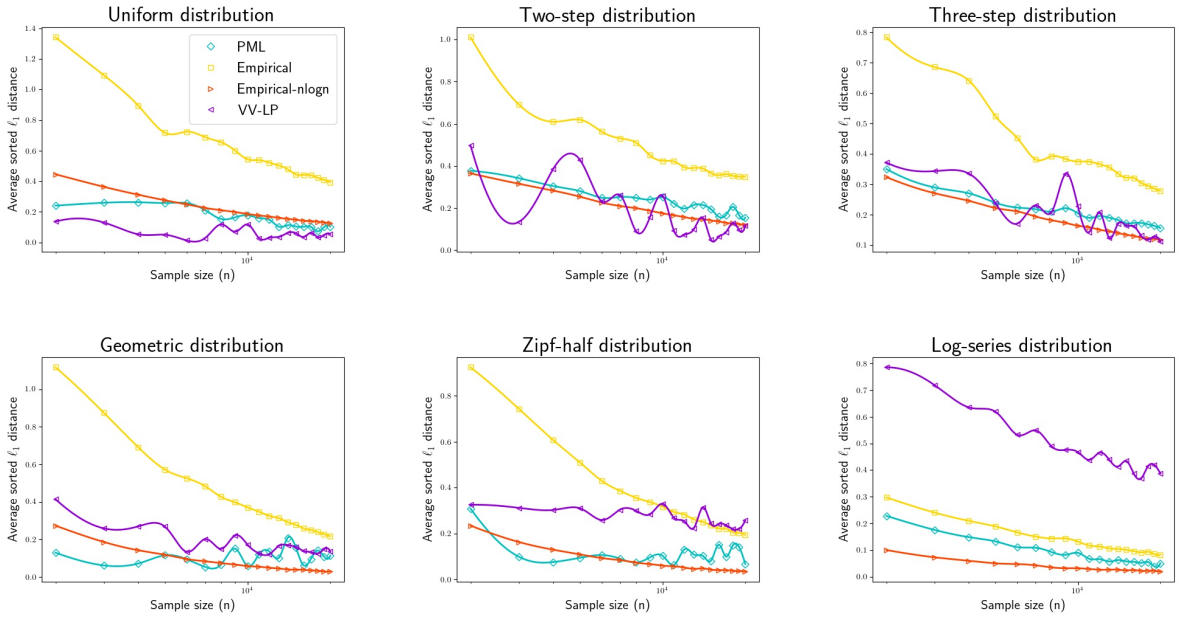


Figure 2.4: Distribution estimation under sorted ℓ_1 distance

Shannon Entropy Estimation under Absolute Error

In Figure 2.5, the sample size n ranges from 1,000 to 1,000,000, and the vertical axis reflects the absolute difference between the true entropy values and the estimates, averaged over 30 independent trials. We compare our estimator with two state-of-the-art estimators, *WY* in [WY16], and *jiao2015minimax* in [JVHW15], as well as the empirical estimator, and the empirical estimator with a larger $n\log n$ sample size. Additional entropy estimators such as the Miller-Mallow estimator [Car69], the best upper bound (BUB) estimator [Pan03], and the Valiant-Valiant estimator [VV11a] were compared in [WY16, JVHW15] and found to perform similarly to or worse than the two estimators that we compared with, therefore we do not include them here. Also, considering [VV11a], page 50 in [Yan16] notes that "the performance of linear programming estimator starts to deteriorate when the sample size is very large."

Note that the alphabet size k is a crucial input to WY, but is not required by either jiao2015minimax or our PML algorithm. In the experiments, we provide WY with the true value of $k = 5{,}000$.

As shown in the plots, our estimator performs as well as these state-of-the-art estimators.

α-Rényi Entropy Estimation under Absolute Error

For a distribution $p \in \Delta_{\mathcal{X}}$, recall that the α-power sum of p is $P_\alpha(p) = \sum_x p(x)^\alpha$, implying $H_\alpha(p) = (1-\alpha)^{-1}\log(P_\alpha(p))$. To establish the sample-complexity upper bounds mentioned in Section 2.3.2 for non-integer α values, [AOST16] first estimate the $P_\alpha(p)$ using the α-power-sum estimator proposed in [JVHW15], and then substitute the estimate into the previous equation. The authors of [JVHW15] have implemented this two-step Rényi entropy estimation algorithm. In the experiments, we take a sample of size n, ranging from 10,000 to 100,000, and compare our estimator with this implementation, referred to as *jiao2015minimax*, the empirical estimator, and the empirical estimator with a larger $n\log n$ sample size. Note that $\log n$ ranges from 9.2 to 11.5. According to the results in [AOST16], the sample complexities for estimating α-Rényi entropy

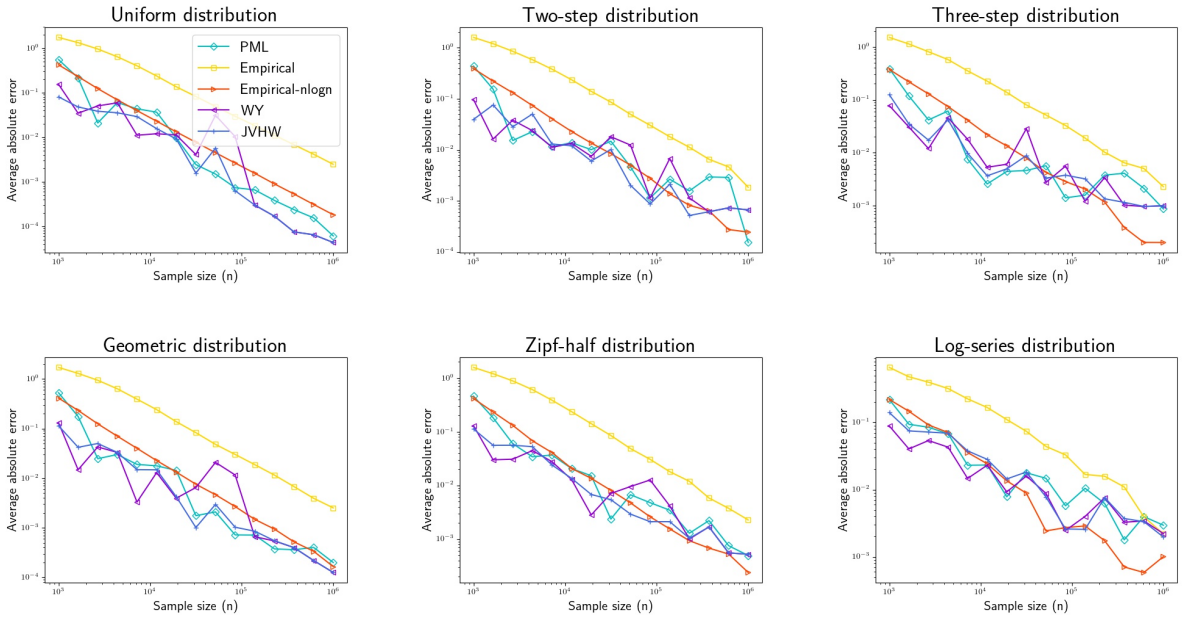


Figure 2.5: Shannon entropy estimation under absolute error

are quite different for $\alpha < 1$ and $\alpha > 1$, hence we consider two cases: $\alpha = 0.5$ and $\alpha = 1.5$.

As shown in Figure 2.6 and 2.7, our estimator clearly outperformed the one proposed by [AOST16, JVHW15].

We further note that for small sample sizes and several distributions, the estimator in [AOST16, JVHW15] performs significantly worse than ours. Also, for large sample sizes, the estimators in [AOST16, JVHW15] degenerates to the simple empirical plug-in estimator. In comparison, our proposed estimator tracks the performance of the empirical estimator with a larger $n \log n$ sample size for nearly all the tested distributions.

2.B Lipschitz-Property Estimation

2.B.1 Proof Outline of Theorem 1

The proof proceeds as follows. First, fixing n, $\mathcal{X}$, and a symmetric additive property f that is 1-Lipschitz on $(\Delta_{\mathcal{X}}, R)$, we consider a related linear program defined in [Val12], and lower

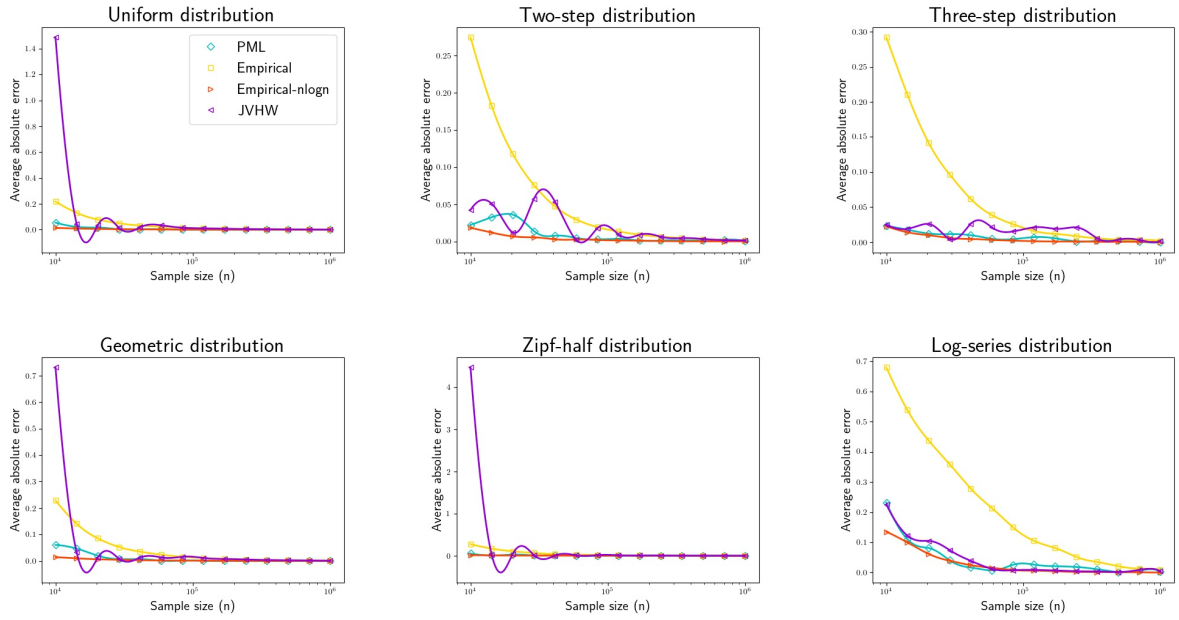


Figure 2.6: 0.5-Rényi entropy estimation under absolute error

bound the worst-case error of any estimators using the linear program's objective value, say v. Second, following the construction in [Val12], we find an explicit estimator $\hat{f}^\star$ that is *linear*, i.e., can be expressed as a linear combination of φ_i's, and show optimality by upper bounding its worst-case error in terms of v. Third, we study the concentration of a general linear estimator, and through the McDiarmid's inequality [McD89], relate the tail probability of its estimate to the estimator's sensitivity to the input changes. Fourth, we bound the sensitivity of $\hat{f}^\star$ by the maximum difference between its consecutive coefficients, and further bound this difference by a function of n, showing that the estimate induced by $\hat{f}^\star$ highly concentrates around its expectation. Finally, we invoke the result in [ADOS17] that the PML-plug-in estimator is competitive to all profile-based estimators whose estimates are highly concentrated, concluding that PML shares the optimality of $\hat{f}^\star$, thereby establishing Theorem 1.

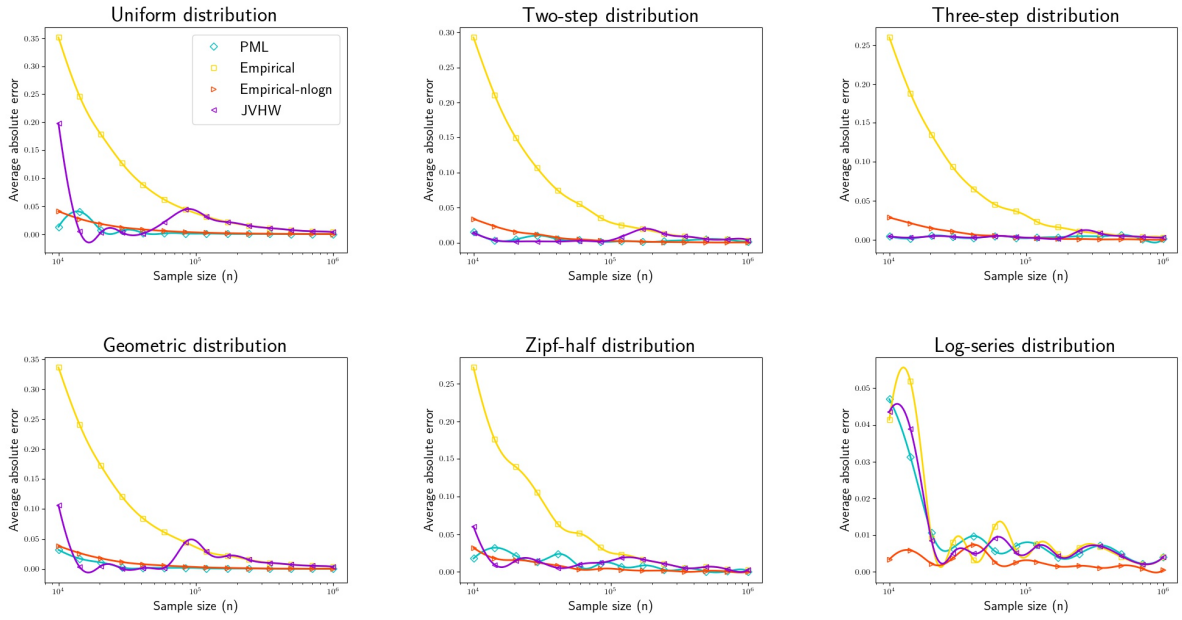


Figure 2.7: 1.5-Rényi entropy estimation under absolute error

2.B.2 Technical Details

Let f be a symmetric additive property that is 1-Lipschitz on $(\Delta_{\mathcal{X}}, R)$. Without loss of generality, we assume that $f(p) = 0$ if $p(x) = 1$ for some $x \in \mathcal{X}$.

Lower bound First, fixing n, $\mathcal{X}$, and f, we lower bound the worst-case error of any estimators.

Let $u \in (0, 1/2)$ be a small absolute constant. If there is an estimator $\hat{f}$ that, when given a length-n sample from any distribution $p \in \Delta_{\mathcal{X}}$, will estimate $f(p)$ up to an error of ε with probability at least $1/2 + u$. Then for any two distributions $p_1, p_2 \in \Delta_{\mathcal{X}}$ satisfying $|f(p_1) - f(p_2)| > \varepsilon$, we can use $\hat{f}$ to distinguish $X^n \sim p_1$ from $X^n \sim p_2$, and will be correct with probability at least $1/2 + u$.

On the other hand, for any parameter $c_1 \in (1/100, 1/25]$ and $c_2 = 1/2 + 6c_1$, consider the corresponding linear program defined in Linear Program 6.7 in [Val12], and denote by v the objective value of any of its solutions. Then, Proposition 6.8 in [Val12] implies that we can find two distributions $p_1, p_2 \in \Delta_{\mathcal{X}}$ such that $|f(p_1) - f(p_2)| > v \cdot (1 - o(1)) - O(n^{-c_1} \log n)$, and no

algorithm can use $\mathrm{Poi}(n)$ sample points to distinguish these two distributions with probability at least $1/2+u$.

The previous reasoning yields that $v < (1+o(1))\varepsilon + O(n^{-c_1}\log n)$. By construction, v is a function of $\mathcal{X}, n$, and f, and essentially serves as a lower bound for ε.

Upper Bound Second, fixing n, $\mathcal{X}$, and f, we construct an explicit estimator based on the previously mentioned linear program, and show optimality by upper bounding its worst-case error in terms of v, the linear program's objective value.

A property estimator $\hat{f}$ is *linear* if there exist real coefficients $\{\ell_i\}_{i\geq 1}$ such that the identity $\hat{f}(x^n) = \sum_{i\geq 1} \ell_i \cdot \varphi_i(x^n)$ holds for all x^n. The following lemma (Proposition 6.10 in [Val12]) bounds the worst-case error of a linear estimator when its coefficients satisfy certain conditions.

Lemma 2. *Given any positive integer m, and real coefficients $\{\beta_i\}_{i\geq 0}$, define $\varepsilon(y) := f(y)/y - e^{-my}\sum_{i\geq 0}\beta_i \cdot (my)^i/i!$. Let $\beta_i^\star := \beta_{i-1}\cdot i/m, \forall i \geq 1$, and $\beta_0^\star := 0$. If for some $a', b', c' > 0$,*

1. *$|\varepsilon(y)| \leq a' + b'/y$,*
2. *$|\beta_j^\star - \beta_\ell^\star| \leq c'\sqrt{j/m}$ for any j and ℓ such that $|j-\ell| \leq \sqrt{j}\log m$,*

then given a sample X^m from any $p \in D_{\mathcal{X}}$, the estimator defined by $\sum_{i\geq 1}\beta_i^\star \cdot \varphi_i$ will estimate $f(p)$ with an accuracy of $a' + b'\cdot k + c'\cdot \log m$ and a failure probability at most $o(1/\mathrm{poly}(m))$.

Following the construction in [Val12] (page 124), let $z := (z_0, z_1, \ldots)$ be the vector of coefficients induced by any solution of the dual program of the previously mentioned linear program. For our purpose, the way in which these coefficients are derived is largely irrelevant. One can show that $|z_\ell| \leq v \cdot n^{c_2}, \forall \ell \geq 0$. Let $t_n := 2n^{-c_1}\log n$ and $\alpha \in (0,1)$, and define

$$\beta_i := (1-e^{-t_n\alpha i}) f\left(\frac{(i+1)\alpha}{n}\right)\frac{n}{(i+1)\alpha} + \sum_{\ell=0}^{i} z_\ell (1-t_n)^\ell \alpha^\ell (1-\alpha)^{i-\ell}\binom{i}{\ell}.$$

for any $i \leq n$, and $\beta_i := \beta_n$ for $i > n$. The next lemma shows that we can find proper parameters a, b, and c to apply Lemma 2 to the above construction. Specifically,

Lemma 3. *For any $\alpha \in [1/100, 1)$ and some $a'', b'' \geq 0$ such that $a'' + b''k \leq v$, if $v \leq \log^2 n$ and c_1, c_2 satisfy $\alpha c_2 + (3/2 - \alpha)c_1 \leq 1/4$, the two conditions in Lemma 2 hold for the above construction with $m = n/\alpha$, $a' = a'' + O(n^{-c_1/2}\log^2 n)$, $b' = b''(1 + O(t_n))$, and $c' = O(n^{-1/4}\log^3 n)$. Furthermore, for any $i \geq 0$, we have $|\beta_i| \leq O(n^{\alpha c_2 + (1-\alpha)c_1}\log^3 n)$.*

This lemma differs from the results established in the proof of Proposition 6.19 in [Val12] only in the applicable range of α, where the latter assumes that $\alpha \in [1/2, 1)$. For completeness, we will present a proof of Lemma 3 in Appendix 2.F.

By Lemma 2 and 3, if $v \leq \log^2 n$, given a sample $X^{n/\alpha}$ from any $p \in \Delta_{\mathcal{X}}$, the linear estimator $\sum_{i \geq 1} \beta_i^\star \cdot \varphi_i$ will estimate $f(p)$ with an accuracy of $a' + b'k + c' \log(n/\alpha) = a'' + O(n^{-c_1/2}\log^2 n) + b''k(1 + O(t_n)) + O(n^{-1/4}\log^4 n) \leq v(1 + O(t_n)) + O(n^{-c_1/2}\log^2 n)$ and a failure probability at most $o(1/\text{poly}(n))$. Recall that for fixed $\mathcal{X}, n$, and f, the value of v is a constant, thus can be computed without samples. Furthermore according to the last claim in Proposition 6.19 in [Val12], for $v > \log^2 n$, the estimator that always returns 0 has an error of at most $(1 + o(1))v$. Hence with high probability, the estimator $\hat{f}^\star := \sum_{i \geq 1} (\beta_i^\star \cdot \mathbb{1}_{v \leq \log^2 n}) \cdot \varphi_i$ will estimate $f(p)$ up to an error of $v(1 + o(1)) + O(t_n \log n)$, for any possible values of v.

Concentration of linear estimators Third, we slightly diverge from the previous discussion and study the concentration of general linear estimators.

The *sensitivity* of a property estimator $\hat{f} : \mathcal{X}^* \to \mathbb{R}$ for a given input size n is

$$s_n(\hat{f}) := \max \{ f(x^n) - f(y^n) : x^n \text{ and } y^n \text{ differ in one element} \},$$

the maximum change in its value when the input sequence is modified at exactly one location. For any $p \in \Delta_{\mathcal{X}}$ and $X^n \sim p$, the following corollary of the McDiarmid's inequality [McD89] relates the two-side tail probability of $\hat{f}(X^n)$ to $s_n(\hat{f})$.

Lemma 4. *For all $t \geq 0$, we have* $\Pr\left(|\hat{f}(X^n) - \mathbb{E}[\hat{f}(X^n)]| \geq t\right) \leq 2\exp(-2t^2 \cdot (\sqrt{n} s_n(\hat{f}))^{-2})$.

Define $\ell_0 := 0$. The next lemma bounds the sensitivity of a linear estimator $\hat{f} := \sum_{i \geq 1} \ell_i \cdot$

φ_i in terms of $\max_{i\geq 1}|\ell_i - \ell_{i-1}|$, the maximum absolute difference between its consecutive coefficients.

Lemma 5. *For any n and linear estimator $\hat{f} := \sum_{i\geq 1}\ell_i \cdot \varphi_i$, we have $s_n(\hat{f}) \leq 2\max_{i\geq 1}|\ell_i - \ell_{i-1}|$.*

Proof. Let x^n and y^n be two arbitrary sequences over $\mathcal{X}$ that differ in one element. Let i be the index where $x_i \neq y_i$. Then by definition, the following multiplicity equalities hold: $\mu_{x_i}(x^n) = \mu_{x_i}(y^n)+1$, $\mu_{y_i}(y^n) = \mu_{y_i}(x^n)+1$, and $\mu_x(x^n) = \mu_x(y^n)$ for $x \in \mathcal{X}$ satisfying $x \neq x_i, y_i$. For simplicity of notation, let $\mu_0 := \mu_{x_i}(x^n)$, $\mu_1 := \mu_{y_i}(y^n)$, and for any $i \geq 1$, let $\hat{f}_i := \ell_{i-1}\cdot\varphi_{i-1} + \ell_i \cdot \varphi_i$.

The first multiplicity equality implies $\varphi_{\mu_0}(x^n) = \varphi_{\mu_0}(y^n)+1$ and $\varphi_{\mu_0-1}(x^n) = \varphi_{\mu_0-1}(y^n) - 1$. Therefore, we have $\hat{f}_{\mu_0}(x^n) - \hat{f}_{\mu_0}(y^n) = \ell_{\mu_0} - \ell_{\mu_0-1}$. Similarly, the second equality implies $\hat{f}_{\mu_1}(x^n) - \hat{f}_{\mu_1}(y^n) = -\ell_{\mu_1} + \ell_{\mu_1-1}$. The third equality combines these two results and yields

$$\hat{f}(x^n) - \hat{f}(y^n) = \ell_{\mu_0} - \ell_{\mu_0-1} + (-\ell_{\mu_1} + \ell_{\mu_1-1}).$$

Applying the triangle inequality to the right-hand side completes the proof. □

By these lemmas, we have the following result for the concentration of linear estimators.

Corollary 1. *For any $t \geq 0$, $p \in \Delta_{\mathcal{X}}$, and $\hat{f} := \sum_{i\geq 1}\ell_i\cdot\varphi_i$, if $X^n \sim p$, then*

$$\Pr\left(|\hat{f}(X^n) - \mathbb{E}[\hat{f}(X^n)]| \geq t\right) \leq 2\min_{i\geq 1}\exp(-t^2\cdot(\sqrt{2n}(\ell_i - \ell_{i-1}))^{-2}).$$

Sensitivity bound Fourth, we bound the sensitivity of $\hat{f}^\star = \sum_{i\geq 1}(\beta_i^\star \cdot \mathbb{1}_{v\leq \log^2 n})\cdot\varphi_i$. By Lemma 5, it suffices to consider the absolute difference between consecutive $\beta_i^\star$'s. We assume $v \leq \log^2 n$ and $\alpha \in [1/100, 1)$, and analyze two cases below, depending on whether i is greater than $400n^{c_1}$ or not. By Lemma 3, for $i \leq 400n^{c_1}$, we have $|\beta_i| \leq O(n^{\alpha c_2 + (1-\alpha)c_1}\log^3 n)$. Define $\beta_{-1} := 0$. Then,

$$|\beta_{i+1}^\star - \beta_i^\star| \leq \left|\frac{400n^{c_1}+1}{n/\alpha}\beta_i\right| + \left|\frac{400n^{c_1}}{n/\alpha}\beta_{i-1}\right| \leq O\left(n^{\alpha c_2 + (2-\alpha)c_1 - 1}\log^3 n\right).$$

For $i > 400n^{c_1}$, we only need to consider $i < n$ since $\beta^\star_{i+1} = \beta^\star_i$ for all $i \geq n$. Then,

$$
\begin{aligned}
|\beta^\star_{i+1} - \beta^\star_i| &\overset{(a)}{\leq} \left| \sum_{\ell=0}^{i} z_\ell (1-t_n)^\ell \alpha^\ell (1-\alpha)^{i-\ell} \binom{i}{\ell} \frac{(i+1)\alpha}{n} \right| \\
&\quad + \left| \sum_{\ell=0}^{i-1} z_\ell (1-t_n)^\ell \alpha^\ell (1-\alpha)^{i-1-\ell} \binom{i-1}{\ell} \frac{i\alpha}{n} \right| \\
&\quad + \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right) \right| + \left| e^{-t_n \alpha i} f\left(\frac{(i+1)\alpha}{n}\right) \right| + \left| e^{-t_n \alpha (i-1)} f\left(\frac{i\alpha}{n}\right) \right| \\
&\overset{(b)}{\leq} (n^{c_2} \log^2 n) \left| \sum_{\ell=0}^{i} (1-t_n)^\ell \alpha^\ell (1-\alpha)^{i-\ell} \binom{i}{\ell} \right| \\
&\quad + (n^{c_2} \log^2 n) \left| \sum_{\ell=0}^{i-1} (1-t_n)^\ell \alpha^\ell (1-\alpha)^{i-1-\ell} \binom{i-1}{\ell} \right| \\
&\quad + \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right) \right| + \left| e^{-t_n \alpha i} f\left(\frac{(i+1)\alpha}{n}\right) \right| + \left| e^{-t_n \alpha (i-1)} f\left(\frac{i\alpha}{n}\right) \right| \\
&\overset{(c)}{\leq} (n^{c_2} \log^2 n)(1-t_n\alpha)^{i-1}(2-t_n\alpha) + \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right) \right| + 2e^{-t_n\alpha(i-1)-1} \\
&\overset{(d)}{\leq} 2(n^{c_2} \log^2 n) \left(1 - \frac{\log n}{50n^{c_1}}\right)^{400n^{c_1}} + \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right) \right| + 2n^{-2}/e \\
&\overset{(e)}{=} 2(n^{c_2} \log^2 n) \left(\left(1 - \frac{\log n}{50n^{c_1}}\right)^{\frac{50n^{c_1}}{\log n}} \right)^{8\log n} + \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right) \right| + 2n^{-2}/e \\
&\overset{(f)}{\leq} 2n^{-2} + \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right) \right|,
\end{aligned}
$$

where (a) follows from the triangle inequality; (b) follows from $i \leq n$, $v \leq \log^2 n$, and $|z_\ell| \leq v \cdot n^{c_2}$ for all $\ell \geq 0$; (c) follows from the binomial theorem and $|f(x)| \leq x|\log x| \leq 1/e$ for $x \in (0,1]$; (d) follows from $\alpha \geq 1/100$, $i > 400n^{c_1}$, and $t_n = 2n^{-c_1}\log n$; (e) follows from simple algebra; and (f) follows from $c_2 = 1/2 + 6c_1 < 1$ and $(1-1/x)^x \leq e^{-1}$ for $x > 1$.

It remains to analyze the second term on the right-hand side.

$$\begin{aligned}
\mathrm{RHS}_2 &:= \left| f\left(\frac{(i+1)\alpha}{n}\right) - f\left(\frac{i\alpha}{n}\right)\right| \\
&\overset{(a)}{=} \frac{(i+1)\alpha}{n}\left| f\left(\frac{(i+1)\alpha}{n}\right)\frac{n}{(i+1)\alpha} - f\left(\frac{i\alpha}{n}\right)\frac{n}{(i+1)\alpha}\right| \\
&\overset{(b)}{=} \frac{(i+1)\alpha}{n}\left| f\left(\frac{(i+1)\alpha}{n}\right)\frac{n}{(i+1)\alpha} - f\left(\frac{i\alpha}{n}\right)\frac{n}{i\alpha} + f\left(\frac{i\alpha}{n}\right)\frac{n}{i(i+1)\alpha}\right| \\
&\overset{(c)}{\le} \frac{(i+1)\alpha}{n}\left|\log\frac{i+1}{i}\right| + \frac{(i+1)\alpha}{n}\left|\frac{i\alpha}{n}\left(\log\left(\frac{i\alpha}{n}\right)\right)\frac{n}{i(i+1)\alpha}\right| \\
&\overset{(d)}{\le} \frac{(i+1)\alpha}{n}\frac{1}{i} + O\left(\frac{\log n}{n}\right) \overset{(e)}{\le} O\left(\frac{\log n}{n}\right),
\end{aligned}$$

where (a), (b) and (e) follows from simple algebra; (c) follows from $|f(x)/x - f(y)/y| \le |\log(x/y)|$ for all $x, y \in (0,1]$; (d) follows from $\log(1+x) \le x$ for $x \ge 0$ and $x|\log x| \le 1/e$ for $x \in (0,1]$.

Consolidating the above inequalities and applying Lemma 5, we get the sensitivity bound

$$s_n(f^\star) \le O\left(n^{\alpha c_2 + (2-\alpha)c_1 - 1}\log^3 n\right).$$

Competitiveness of PML A property estimator $\hat{f}$ is *profile-based* if there exists a mapping $\hat{g}$ such that $\hat{f}(x^n) = \hat{g}(\varphi(x^n))$ for all $x^n \in \mathcal{X}^*$. The following lemma [ADJ+12a, ADOS17, Das12] states that the PML estimator is competitive to other profile-based estimators.

Lemma 6. *For any positive real numbers* ε *and* δ*, additive symmetric property* f*, and profile-based estimator* $\hat{f}$*, the PML-plug-in estimator* $f(p_\varphi)$ *satisfies*

$$n_f(f(p_\varphi), 2\varepsilon, \delta\cdot\exp(3\sqrt{n})) \le n_f(\hat{f}, \varepsilon, \delta).$$

For any β*-approximate PML, a similar result holds with* $\delta\cdot\exp(3\sqrt{n})$ *replaced by* $\delta\cdot\exp(3\sqrt{n})/\beta$.

The factor $\exp(3\sqrt{n})$ directly comes from the well-known result of [HR18] on integer

partitions, since there is a bijective mapping from profiles of size n to partitions of integer n.

Final analysis Finally, we combine the above results and establish Theorem 1.

Denote by $\tau(n)$ the previous upper bound on $s_n(f^\star)$. Let p be a distribution in $\Delta_\mathcal{X}$ and $X^n \sim p$. Let γ be an absolute constant in $(0, 1/4)$. Then by Lemma 4,

$$\Pr\left(|\hat{f}^\star(X^n) - \mathbb{E}[\hat{f}^\star(X^n)]| \geq 2n^{1-\gamma}\tau(n)\right) \leq 2\exp(-8n^{1-2\gamma}).$$

Let $\varepsilon > 0$ be an error parameter. Assume there exists an estimator $\hat{f}$ that, when given a length-αn sample from any distribution $p' \in \Delta_\mathcal{X}$, estimates $f(p')$ up to an absolute error ε with probability at least $2/3$. Then according to the results in the upper- and lower-bound sections, with probability at most $o(1/\mathrm{poly}(n))$, the estimate $\hat{f}^\star(X^n)$ will differ from $f(p)$ by more than $v(1+o(1)) + O(n^{-c_1/2}\log^2 n) \leq \varepsilon(1+o(1)) + O(n^{-c_1/2}\log^2 n)$. In addition, by the equality $\sum_{i\geq 1} i \cdot \varphi_i(X^n) = n$ and Lemma 3, we surely have $|\hat{f}^\star(X^n)| \leq |\sum_{i\geq 1}(i/m)\beta_{i-1} \cdot \varphi_i(X^n)| \leq \max_{i\geq 0}|\beta_i| \leq O(n^{\alpha c_2 + (1-\alpha)c_1}\log^3 n)$. Multiplying this bound by $o(1/\mathrm{poly}(n))$ yields a quantity that is negligible comparing to $O(n^{-c_1/2}\log^2 n)$. Therefore, the absolute bias $|\mathbb{E}[\hat{f}^\star(X^n)] - f(p)|$ is at most $\varepsilon(1+o(1)) + O(n^{-c_1/2}\log^2 n)$. The triangle inequality combines this with the tail bound above:

$$\Pr\left(|\hat{f}(X^n) - f(p)| \geq \varepsilon\,(1+o(1)) + O(n^{-c_1/2}\log^2 n) + 2n^{1-\gamma}\tau(n)\right) \leq 2\exp\left(-8n^{1-2\gamma}\right).$$

Let $\alpha = 1/4$. For PML and APML estimators, set (γ, c_1) to be $(1/4, 1/31)$ and $(0.166, 1/91)$, respectively. Combined, the last inequality and Lemma 6 imply Theorem 1. There is a simple trade-off between α and c_1 induced by our proof technique. Specifically, if we increase the value of c_1 to achieve a better lower bound on ε, the value of α may need to be reduced accordingly, which enlarges the sample complexity gap between our estimators and the optimal one. For example, reducing α to $1/12$ and $1/22$, we can improve c_1 to $1/25$ and $1/20$, respectively, for both PML and APML.

2.C α-Rényi Entropy Estimation

For any $p \in \Delta_{\mathcal{X}}$ and non-negative $\alpha \neq 1$, the α-*Rényi entropy* [Rén61] of p is

$$H_\alpha(p) := \frac{1}{1-\alpha} \log P_\alpha(p) = \frac{1}{1-\alpha} \log \left(\sum_x p(x)^\alpha \right).$$

For $\mathcal{X}$ of finite size k and any $p \in \Delta_{\mathcal{X}}$, it is well-known that $H_\alpha(p) \in [0, \log k]$.

2.C.1 Proof of Theorem 2: $\alpha \in (3/4, 1)$

For $\alpha \in (3/4, 1)$, the following theorem characterizes the performance of the PML-plug-in estimator. For any distribution $p \in \Delta_{\mathcal{X}}$, error parameter $\varepsilon \in (0,1)$, and sampling parameter n, draw a sample $X^n \sim p$ and denote its profile by φ. Then for sufficiently large k,

Theorem 2. *For an* $\alpha \in (3/4, 1)$*, if* $n = \Omega_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha} \log k))$*,*

$$\Pr\left(|H_\alpha(p_\varphi) - H_\alpha(p)| \geq \varepsilon \right) \leq \exp(-\sqrt{n}).$$

We establish both this theorem and an analogous result for APML in the remaining section. Let n be a sampling parameter and $p \in \Delta_{\mathcal{X}}$ be an unknown distribution. For some α-dependent positive constants $c_{\alpha,1}$ and $c_{\alpha,2}$ to be determined later, let $\tau := c_{\alpha,1} \log n$ and $d := c_{\alpha,2} \log n$ be threshold and degree parameters, respectively. Let N, N' be independent Poisson random variables with mean n. Consider Poisson sampling with two samples drawn from p, first of size N and the second N'. Suppressing the sample representations, for each $x \in \mathcal{X}$, we denote by μ_x and μ'_x the multiplicities of symbol x in the first and second samples, respectively. Denote by $q(z) := \sum_{m=0}^d a_m z^m$ be the degree-d min-max polynomial approximation of z^a over $[0,1]$. We consider the following variant of the polynomial-based estimator proposed in [AOST16].

$$\hat{P}_\alpha := \sum_x \left(\sum_{m=0}^d \frac{a_m (2\tau)^{\alpha-m} \mu_x^{\underline{m}}}{n^\alpha} \right) \mathbb{1}_{\mu_x \leq 4\tau} \cdot \mathbb{1}_{\mu'_x \leq \tau} + \sum_x \left(\frac{\mu_x}{n} \right)^\alpha \mathbb{1}_{\mu'_x > \tau}.$$

The smaller the value of μ'_x is, the smaller we expect the value of $p(x)$ to be. In view of this, we denote the first and second components of $\hat{P}_\alpha$ by $\hat{P}^{(s)}_\alpha$ and $\hat{P}^{(\ell)}_\alpha$, and refer to them as small- and large-probability estimators, respectively. Note that our estimator differs from that in [AOST16] only by the additional $\mathbb{1}_{\mu_x \le 4\tau}$ term, which for sufficiently large $c_{\alpha,1}$, only modifies $\mathbb{E}[\hat{P}^{(s)}_\alpha]$ by at most $n^{-2\alpha}$.

Note that μ' naturally induces a partition over $\mathcal{X}$. For symbols x with $\mu_x \le 4\tau$, we denote

$$P^{(s)}_{a,\mu'}(p) := \sum_{x:\mu_x \le 4\tau} p(x)^\alpha,$$

the small-probability power sum. Analogously, for symbols x with $\mu_x > 4\tau$, we denote by

$$P^{(\ell)}_{a,\mu'}(p) := \sum_{x:\mu_x > 4\tau} p(x)^\alpha$$

the large-probability power sum. These are random properties with non-trivial variances and are hard to be analyzed. To address this, we apply an "expectation trick" and denote by $P^{(s)}_a(p) := \mathbb{E}[P^{(s)}_{a,\mu'}(p)]$ and $P^{(\ell)}_a(p) := \mathbb{E}[P^{(\ell)}_{a,\mu'}(p)]$ their expected values, both of which are additive symmetric properties.

Let ε be a given error parameter and $n = \Omega_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha}\log k))$ be a sampling parameter. First we consider the small probability estimator. By the results in [AOST16], for sufficiently large $c_{\alpha,1}$, the bias of $\hat{P}^{(s)}_\alpha$ in estimating $P^{(s)}_\alpha(p)$ satisfies

$$|\mathbb{E}[\hat{P}^{(s)}_\alpha] - P^{(s)}_\alpha(p)| \le O_\alpha(1) \cdot P_\alpha(p) \left(\frac{k}{n\log n}\right)^\alpha + n^{-\alpha} \le \varepsilon P_\alpha(p),$$

where we have used $n^{-\alpha} = O_\alpha(\varepsilon k^{-1}(\log k)^\alpha) \le \varepsilon P_\alpha(p)$. To show concentration, we bound the sensitivity of estimator $\hat{P}^{(s)}_\alpha$. For $m \ge 0$, we can bound the coefficients of $q(x)$ as follows.

$$|a_m| = O_\alpha((\sqrt{2}+1)^d) = O_\alpha(n^{c_{\alpha,2}}).$$

Therefore by definition, changing one point in the sample changes the value of $\hat{P}_\alpha^{(s)}$ by at most

$$2\left(\sum_{m=0}^{d}\frac{|a_m|(2\tau)^{\alpha-m}(4\tau)^{\underline{m}}}{n^\alpha}\right)\le\sum_{m=0}^{d}\frac{|a_m|(2\tau)^\alpha 2^{m+1}}{n^\alpha}=O_\alpha\left(n^{2c_{\alpha,2}-\alpha}(\log n)^\alpha\right).$$

Let $\lambda\in(0,1/4)$ be an arbitrary absolute constant. For sufficiently small $c_{\alpha,2}$, the right-hand side is at most $O_\alpha\left(n^{\lambda-\alpha}\right)$. The McDiarmid's inequality together with the concentration of Poisson random variables implies that for all $\varepsilon\ge 0$,

$$\Pr\left(|\hat{P}_\alpha^{(s)}-\mathbb{E}[\hat{P}_\alpha^{(s)}]|\ge\varepsilon P_\alpha(p)\right)\le 2\exp(-\Omega_\alpha(\varepsilon^2P_\alpha^2(p)n^{2\alpha-1-2\lambda})).$$

Note that $n=\Omega_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha}\log k))$ and $P_\alpha(p)\ge 1$, which follows from the fact that z^α is a concave function over $[0,1]$ for $\alpha\in(0,1)$. Hence we obtain

$$\Pr\left(|\hat{P}_\alpha^{(s)}-\mathbb{E}[\hat{P}_\alpha^{(s)}]|\ge\varepsilon P_\alpha(p)\right)\le 3\exp\left(-\Omega_\alpha\left(\varepsilon^2n^{2\alpha-1-2\lambda}\right)\right).$$

For $\alpha>3/4$, we can set $\lambda=(4\alpha-3)/8$. Direct calculation shows that for sufficiently large k, the right-hand side is no more than $\exp(-8\sqrt{n})$. Analogously, we can show that for $\alpha>5/6$, the probability bound can be improved to $\exp(-\Theta(n^{2/3}))$.

Second, we consider the large probability estimator. To begin with, we set $n=\Theta_\alpha(k^{1/3})$. By the results in [AOST16], for sufficiently large $c_{\alpha,1}$, the bias of $\hat{P}_\alpha^{(\ell)}$ in estimating $P_\alpha^{(\ell)}(p)$ satisfies

$$|\mathbb{E}[\hat{P}_\alpha^{(\ell)}]-P_\alpha^{(\ell)}(p)|\le O_\alpha\left(\frac{P_\alpha(p)}{\tau}\right)+\frac{1}{n^\alpha},$$

which, for sufficiently large k, is at most $\varepsilon P_\alpha(p)$. Under the same conditions, the variance of $\hat{P}_\alpha^{(\ell)}$ is at most

$$\mathrm{Var}(\hat{P}_\alpha^{(\ell)})\le O_\alpha\left(\sum_x\frac{p(x)^{2\alpha}}{\tau}\right)+\frac{1}{n^{2\alpha}}\le\frac{(\varepsilon P_\alpha(p))^2}{3}.$$

Then, the Chebyshev's inequality yields

$$\Pr\left(|\mathbb{E}[\hat{P}_\alpha^{(\ell)}] - \hat{P}_\alpha^{(\ell)}| \geq \varepsilon P_\alpha(p)\right) \leq \frac{1}{3}.$$

The triangle inequality combines this tail bound with the above bias bound and implies

$$\Pr\left(|P_\alpha^{(\ell)}(p) - \hat{P}_\alpha^{(\ell)}| \geq 2\varepsilon P_\alpha(p)\right) \leq \frac{1}{3}.$$

Therefore, utilizing the median trick and $\alpha < 1$, we can construct another estimator $\hat{P}_\alpha^{(\ell,1)}$ that takes a sample of size $n = \Omega_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha}\log k))$, and satisfies

$$\Pr\left(|P_\alpha^{(\ell)}(p) - \hat{P}_\alpha^{(\ell,1)}| \geq 2\varepsilon P_\alpha(p)\right) \leq 2\exp(-\Omega_\alpha(n/k^{1/3}))) \leq 2\exp(-\Theta(n^{2/3})).$$

Recall that $P_\alpha(p) = P_\alpha^{(s)}(p) + P_\alpha^{(\ell)}(p)$. By the union bound and the triangle inequality, under Poisson sampling with parameter $n = \Theta_\alpha(k^{1/\alpha}/(\varepsilon^{1/\alpha}\log k))$,

$$\Pr\left(|P_\alpha(p) - (\hat{P}_\alpha^{(s)} + \hat{P}_\alpha^{(\ell,1)})| \geq 4\varepsilon P_\alpha(p)\right) \leq \exp(-8\sqrt{n}).$$

Since both N and N' are Poisson random variables with mean n, we must have $N + N' \sim \mathrm{Poi}(2n)$, implying that $\Pr(N + N' = 2n) = e^{-2n}(2n)^{2n}/(2n)!$. A variant of the well-known Stirling's formula states that $m! \geq em^{m+1/2}e^{-m}$ for all positive integers m. We obtain $\Pr(N + N' = 2n) \geq e^{-2n}(2n)^{2n} \cdot (e(2n)^{2n+1/2}e^{-2n})^{-1} \geq 1/(e\sqrt{2n}) > 1/(4n)$. Hence, under fixed sampling with a sample size of $2n$, the estimator $\hat{P}_\alpha^{(1)} := (\hat{P}_\alpha^{(s)} + \hat{P}_\alpha^{(\ell,1)})$ satisfies

$$\Pr\left(|P_\alpha(p) - \hat{P}_\alpha^{(1)}| \geq 4\varepsilon P_\alpha(p)\right) \leq 4n\exp(-8\sqrt{n}).$$

Replacing n with $n/2$ and ε with $\varepsilon/4$, the *sufficiency of profiles* [AOST16] implies the existence

of a profile-based estimator $\hat{P}_\alpha^\star$ such that for any $p \in \Delta_\mathcal{X}$,

$$\Pr_{X^n \sim p}\left(|P_\alpha(p) - \hat{P}_\alpha^\star(X^n)| \geq \varepsilon P_\alpha(p)\right) \leq 2n\exp(-4\sqrt{2n}) < \exp(-4\sqrt{n}).$$

Let δ denote the quantity on the right-hand side. For any x^n with profile φ satisfying both $p(\varphi) > \delta$, we must have $|\hat{P}_\alpha^\star(x^n) - P_\alpha(p)| \leq \varepsilon P_\alpha(p)$. By definition, we also have $p_\varphi(\varphi) \geq p(\varphi) > \delta$ and hence $|\hat{P}_\alpha^\star(x^n) - P_\alpha(p_\varphi)| \leq \varepsilon P_\alpha(p_\varphi)$. For any $\varepsilon \in (0, 1/2)$, simple algebra combines the two property inequalities and yields

$$|P_\alpha(p) - P_\alpha(p_\varphi)| \leq 2\varepsilon P_\alpha(p).$$

On the other hand, for a sample $X^n \sim p$ with profile φ', the probability that we have $p(\varphi') \leq \delta$ is at most δ times the cardinality of the set $\Phi^n := \{\varphi(x^n) : x^n \in \mathcal{X}^n\}$. The latter quantity corresponds to the number of integer partitions of n, which, by the well-known result of [HR18], is at most $\exp(3\sqrt{n})$. Hence, the probability that $p(\varphi') \leq \delta$ is upper bounded by $\exp(-\sqrt{n})$. To conclude, we have shown that

$$\Pr\left(|P_\alpha(p) - P_\alpha(p_\varphi)| \geq 2\varepsilon P_\alpha(p)\right) \leq \exp(-\sqrt{n}).$$

In terms of Rényi entropy values, applying the inequality $e^z - 1 \geq 1 - e^{-z} \geq z/2$ for all $z \geq 0$, we establish that for $\alpha > 3/4$ and $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$,

$$\Pr\left(|H_\alpha(p) - H_\alpha(p_\varphi)| \geq \varepsilon\right) = \Pr\left(P_\alpha(p_\varphi)e^{-(\alpha-1)\varepsilon} \leq P_\alpha(p) \leq P_\alpha(p_\varphi)e^{(\alpha-1)\varepsilon}\right) \leq \exp(-\sqrt{n}).$$

2.C.2 Proof of Theorem 3: Non-Integer $\alpha > 1$

The proof of the following theorem is essentially the same as that shown in the previous section. However, for completeness, we still include a full-length proof.

For any distribution $p \in \Delta_{\mathcal{X}}$, error parameter $\varepsilon \in (0,1)$, absolute constant $\lambda \in (0, 0.1)$, and sampling parameter n, draw a sample $X^n \sim p$ and denote its profile by φ. Then for sufficiently large integer k,

Theorem 3. *For a non-integer $\alpha > 1$, if $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$,*

$$\Pr\left(|H_\alpha(p_\varphi) - H_\alpha(p)| \geq \varepsilon\right) \leq \exp(-n^{1-\lambda}).$$

We establish this theorem in the remaining section. Let n be a sampling parameter and $p \in \Delta_{\mathcal{X}}$ be an unknown distribution. For some α-dependent positive constants $c_{\alpha,1}$ and $c_{\alpha,2}$ to be determined later, let $\tau := c_{\alpha,1}\log n$ and $d := c_{\alpha,2}\log n$ be threshold and degree parameters, respectively. Let N, N' be independent Poisson random variables with mean n. Consider Poisson sampling with two samples drawn from p, first of size N and the second N'. Suppressing the sample representations, for each $x \in \mathcal{X}$, we denote by μ_x and μ'_x the multiplicities of symbol x in the first and second samples, respectively. Denote by $q(z) := \sum_{m=0}^{d} a_m z^m$ be the degree-d min-max polynomial approximation of z^a over $[0,1]$. We consider the following variant of the estimator proposed in [AOST16].

$$\hat{P}_\alpha := \sum_x \left(\sum_{m=0}^{d} \frac{a_m (2\tau)^{\alpha-m} \mu_x^{\underline{m}}}{n^\alpha} \right) \mathbb{1}_{\mu_x \leq 4\tau} \cdot \mathbb{1}_{\mu'_x \leq \tau} + \sum_x \left(\frac{\mu_x}{n}\right)^\alpha \mathbb{1}_{\mu'_x > \tau}.$$

The smaller the value of μ'_x is, the smaller we expect the value of $p(x)$ to be. In view of this, we denote the first and second components of $\hat{P}_\alpha$ by $\hat{P}_\alpha^{(s)}$ and $\hat{P}_\alpha^{(\ell)}$, and refer to them as small- and large-probability estimators, respectively. Note that our estimator differs from that in [AOST16] only by the additional $\mathbb{1}_{\mu_y \leq 4\tau}$ term, which for sufficiently large $c_{\alpha,1}$, only modifies $\mathbb{E}[\hat{P}_\alpha^{(s)}]$ by at most $n^{-2\alpha}$.

Note that μ' naturally induces a partition over $\mathcal{X}$. For symbols x with $\mu_x \leq 4\tau$, we denote

by

$$P_{a,\mu'}^{(s)}(p) := \sum_{x:\mu_x \le 4\tau} p(x)^\alpha$$

the small-probability power sum. Analogously, for symbols x with $\mu_x > 4\tau$, we denote by

$$P_{a,\mu'}^{(\ell)}(p) := \sum_{x:\mu_x > 4\tau} p(x)^\alpha$$

the large-probability power sum. These are random properties with non-trivial variances and are hard to be analyzed. To address this, we apply an "expectation trick" and denote by $P_a^{(s)}(p) := \mathbb{E}[P_{a,\mu'}^{(s)}(p)]$ and $P_a^{(\ell)}(p) := \mathbb{E}[P_{a,\mu'}^{(\ell)}(p)]$ their expected values, both of which are additive symmetric properties.

Let ε be a given error parameter and $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$ be a sampling parameter. First we consider the small probability estimator. By the results in [AOST16], for sufficiently large $c_{\alpha,1}$, the bias of $\hat{P}_\alpha^{(s)}$ in estimating $P_\alpha^{(s)}(p)$ satisfies

$$|\mathbb{E}[\hat{P}_\alpha^{(s)}] - P_\alpha^{(s)}(p)| \le O_\alpha(1) \cdot P_\alpha(p) \left(\frac{k}{n\log n}\right)^\alpha + n^{-\alpha} \le \varepsilon P_\alpha(p),$$

where we have used $n^{-\alpha} = O_\alpha(\varepsilon k^{-\alpha}(\log k)^\alpha) \le \varepsilon P_\alpha(p)$. To show concentration, we bound the sensitivity of estimator $\hat{P}_\alpha^{(s)}$. For $m \ge 0$, we can bound the coefficients of $q(x)$ as follows.

$$|a_m| \le O_\alpha((\sqrt{2}+1)^d) = O_\alpha(n^{c_{\alpha,2}}).$$

Therefore by definition, changing one point in the sample changes the value of $\hat{P}_\alpha^{(s)}$ by at most

$$2\left(\sum_{m=0}^{d} \frac{|a_m|(2\tau)^{\alpha-m}(4\tau)^{\underline{m}}}{n^\alpha}\right) \le \sum_{m=0}^{d} \frac{|a_m|(2\tau)^\alpha 2^{m+1}}{n^\alpha} \le O_\alpha\left(n^{2c_{\alpha,2}-\alpha}(\log n)^\alpha\right).$$

Let $\lambda \in (0, 1/4)$ be an arbitrary absolute constant. For sufficiently small $c_{\alpha,2}$, the right-hand side

is at most $O_\alpha\left(n^{\lambda-\alpha}\right)$. The McDiarmid's inequality together with the concentration of Poisson random variables implies that for all $\varepsilon \geq 0$,

$$\Pr\left(|\hat{P}_\alpha^{(s)} - \mathbb{E}[\hat{P}_\alpha^{(s)}]| \geq \varepsilon P_\alpha(p)\right) \leq 2\exp(-\Omega_\alpha(\varepsilon^2 P_\alpha^2(p) n^{2\alpha-1-2\lambda})).$$

Note that $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$ and $P_\alpha(p) \geq k^{1-\alpha}$. Hence we obtain

$$\Pr\left(|\hat{P}_\alpha^{(s)} - \mathbb{E}[\hat{P}_\alpha^{(s)}]| \geq \varepsilon P_\alpha(p)\right) \leq 3\exp\left(-\Omega_\alpha(\varepsilon^2 k^{2-2\alpha} n^{2\alpha-1-2\lambda})\right).$$

By simple algebra, for sufficiently large k, the right-hand side is at most $\exp(-n^{1-3\lambda})$.

Second, we consider the large probability estimator. To begin with, we set $n = \Theta_\alpha(k^\lambda)$. By the results in [AOST16], for sufficiently large $c_{\alpha,1}$, the bias of $\hat{P}_\alpha^{(\ell)}$ in estimating $P_\alpha^{(\ell)}(p)$ satisfies

$$|\mathbb{E}[\hat{P}_\alpha^{(\ell)}] - P_\alpha^{(\ell)}(p)| \leq O_\alpha\left(\frac{P_\alpha(p)}{\tau}\right) + \frac{1}{n^{4\alpha}},$$

which, for sufficiently large k, is at most $\varepsilon P_\alpha(p)$. Under the same conditions, the variance of $\hat{P}_\alpha^{(\ell)}$ is at most

$$\mathrm{Var}(\hat{P}_\alpha^{(\ell)}) \leq O_\alpha\left(\sum_x \frac{p(x)^{2\alpha}}{\tau}\right) + \frac{1}{n^{8\alpha}} \leq \frac{(\varepsilon P_\alpha(p))^2}{3}.$$

Then, the Chebyshev's inequality yields

$$\Pr\left(|\mathbb{E}[\hat{P}_\alpha^{(\ell)}] - \hat{P}_\alpha^{(\ell)}| \geq \varepsilon P_\alpha(p)\right) \leq \frac{1}{3}.$$

The triangle inequality combines this tail bound with the above bias bound and implies

$$\Pr\left(|P_\alpha^{(\ell)}(p) - \hat{P}_\alpha^{(\ell)}| \geq 2\varepsilon P_\alpha(p)\right) \leq \frac{1}{3}.$$

Therefore, utilizing the median trick, we can construct another estimator $\hat{P}_\alpha^{(\ell,1)}$ that takes a sample

of size $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$, and for sufficiently large k, satisfies

$$\Pr\left(|P_\alpha^{(\ell)}(p) - \hat{P}_\alpha^{(\ell,1)}| \geq 2\varepsilon P_\alpha(p)\right) \leq 2\exp(-\Omega_\alpha(n/k^\lambda)) \leq \exp(-n^{1-2\lambda}).$$

Recall that $P_\alpha(p) = P_\alpha^{(s)}(p) + P_\alpha^{(\ell)}(p)$. By the union bound and the triangle inequality, under Poisson sampling with parameter $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$,

$$\Pr\left(|P_\alpha(p) - (\hat{P}_\alpha^{(s)} + \hat{P}_\alpha^{(\ell,1)})| \geq 4\varepsilon P_\alpha(p)\right) \leq \exp(-n^{1-3\lambda}).$$

Since both N and N' are Poisson random variables with mean n, we must have $N + N' \sim \mathrm{Poi}(2n)$, implying that $\Pr(N + N' = 2n) = e^{-2n}(2n)^{2n}/(2n)!$. A variant of the well-known Stirling's formula states that $m! \geq em^{m+1/2}e^{-m}$ for all positive integers m. We obtain $\Pr(N + N' = 2n) \geq e^{-2n}(2n)^{2n} \cdot (e(2n)^{2n+1/2}e^{-2n})^{-1} \geq 1/(e\sqrt{2n}) > 1/(4n)$. Hence, under fixed sampling with a sample size of $2n$, the estimator $\hat{P}_\alpha^{(1)} := (\hat{P}_\alpha^{(s)} + \hat{P}_\alpha^{(\ell,1)})$ satisfies

$$\Pr\left(|P_\alpha(p) - \hat{P}_\alpha^{(1)}| \geq 4\varepsilon P_\alpha(p)\right) \leq 4n\exp(-n^{1-3\lambda}).$$

Replacing ε with $\varepsilon/4$ and λ with $\lambda/5$, the sufficiency of profiles implies the existence of a profile-based estimator $\hat{P}_\alpha^\star$ such that for sufficiently large k and any $p \in \Delta_\mathcal{X}$,

$$\Pr_{X^n \sim p}\left(|P_\alpha(p) - \hat{P}_\alpha^\star(X^n)| \geq \varepsilon P_\alpha(p)\right) \leq 4n\exp(-n^{1-3\lambda/5}) < \exp(-n^{1-4\lambda/5}).$$

Let δ denote the quantity on the right-hand side. For any x^n with profile φ satisfying both $p(\varphi) > \delta$, we must have $|\hat{P}_\alpha^\star(x^n) - P_\alpha(p)| \leq \varepsilon P_\alpha(p)$. By definition, we also have $p_\varphi(\varphi) \geq p(\varphi) > \delta$ and hence $|\hat{P}_\alpha^\star(x^n) - P_\alpha(p_\varphi)| \leq \varepsilon P_\alpha(p_\varphi)$. For any $\varepsilon \in (0, 1/2)$, simple algebra combines the two property inequalities and yields

$$|P_\alpha(p) - P_\alpha(p_\varphi)| \leq 2\varepsilon P_\alpha(p).$$

On the other hand, for a sample $X^n \sim p$ with profile φ', the probability that we have $p(\varphi') \le \delta$ is at most δ times the cardinality of the set $\Phi^n := \{\varphi(x^n) : x^n \in \mathcal{X}^n\}$. The latter quantity corresponds to the number of integer partitions of n, which, by the well-known result of [HR18], is at most $\exp(3\sqrt{n})$. Hence, the probability that $p(\varphi') \le \delta$ is upper bounded by $\exp(-n^{1-\lambda})$. To conclude, we have shown that

$$\Pr\left(|P_\alpha(p) - P_\alpha(p_\varphi)| \ge 2\varepsilon P_\alpha(p)\right) \le \exp(-n^{1-\lambda}).$$

In terms of Rényi entropy values, applying the inequality $e^z - 1 \ge 1 - e^{-z} \ge z/2$ for all $z \ge 0$, we establish that for $n = \Omega_\alpha(k/(\varepsilon^{1/\alpha}\log k))$,

$$\Pr\left(|H_\alpha(p) - H_\alpha(p_\varphi)| \ge \varepsilon\right) = \Pr\left(P_\alpha(p_\varphi)e^{-(\alpha-1)\varepsilon} \le P_\alpha(p) \le P_\alpha(p_\varphi)e^{(\alpha-1)\varepsilon}\right) \le \exp(-n^{1-\lambda}).$$

2.C.3 Proof of Theorem 4: Integer $\alpha > 1$

For an integer $\alpha > 1$, the following theorem characterizes the performance of the PML-plug-in estimator. For any $p \in \Delta_\mathcal{X}$, $\varepsilon \in (0,1)$, and a sample $X^n \sim p$ with profile φ,

Theorem 4. *If* $n = \Omega_\alpha(k^{1-1/\alpha}(\varepsilon^2|\log\varepsilon|)^{-(1+\alpha)})$ *and* $H_\alpha(p) \le (\log n)/4$,

$$\Pr(|H_\alpha(p_\varphi) - H_\alpha(p)| \ge \varepsilon) \le 1/3.$$

Due to the lower bounds in [AOST16], for all possible values of α, the sample complexity of the PML plug-in estimator has the optimal dependency in k. The remaining section is devoted to proving the above theorem. Note that estimating the Rényi entropy $H_\alpha(p)$ to an additive error is equivalent to estimating the power sum $P_\alpha(p)$ to a corresponding multiplicative error. Given

this fact, we consider the estimator $\hat{P}_\alpha$ in [AOST16] that maps each sequence $x^n \in \mathcal{X}^*$ to

$$\hat{P}_\alpha(x^n) := \sum_x \frac{\mu_x(x^n)^{\underline{\alpha}}}{n^{\underline{\alpha}}},$$

where for any real number z, the expression $z^{\underline{\alpha}}$ denotes the falling factorial of z to the power α. For a sample $X^n \sim p$, we have $\mathbb{E}[\hat{P}_\alpha(X^n)] = P_\alpha(p)$. The following lemma [OS17, AOST16] states that $\hat{P}_\alpha(X^n)$ often estimates $P_\alpha(p)$ to a small multiplicative error when n is large.

Lemma 7. *Under the above conditions, for any $\varepsilon, n > 0$,*

$$\Pr\left(|\hat{P}_\alpha(X^n) - P_\alpha(p)| \geq \varepsilon P_\alpha(p)\right) = O_\alpha(\varepsilon^{-2} n^{-1} (P_\alpha(p))^{-1/\alpha}).$$

For sufficiently large $n = \Omega_\alpha(k^{(\alpha-1)/\alpha})$, this inequality together with $P_\alpha(p) \leq k^{1-\alpha}$ implies that

$$\Pr\left(|\hat{P}_\alpha(X^n) - P_\alpha(p)| \geq \frac{1}{2} \cdot P_\alpha(p)\right) \leq \frac{1}{4}.$$

The following corollary is a consequence of the above lemma, the sufficiency of profiles, and the standard median trick.

Corollary 2. *Under the above conditions, there is an estimator $\hat{P}^\star_\alpha$ such that for any $\varepsilon, n > 0$,*

$$\Pr\left(|\hat{P}^\star_\alpha(X^n) - P_\alpha(p)| \geq \varepsilon P_\alpha(p)\right) \leq 2\exp\left(-\Omega_\alpha(\varepsilon^2 n (P_\alpha(p))^{1/\alpha})\right).$$

In addition, the estimator $\hat{P}^\star_\alpha$ is profile-based.

For simplicity, suppress X^n in $p_\mu(X^n)$. Since the profile probability $p(\varphi)$ is invariant to symbol permutation, for our purpose, we can assume that $p_\mu(y) \leq p_\mu(z)$ iff $p_\varphi(x) \leq p_\varphi(y)$, for all $x, y \in \mathcal{X}$. Under this assumption, the following lemma [OSVZ11, AGZ17] relates p_φ to p_μ.

Lemma 8. *For a distribution p and sample $X^n \sim p$ with profile φ,*

$$\Pr\left(\max_x |p_\varphi(x) - p_\mu(x)| > \frac{2\log n}{n^{1/4}}\right) = O\left(\frac{1}{n}\right).$$

Consider $\varepsilon \in (0, 1/2)$ and x^n satisfying $|\hat{P}^\star_\alpha(x^n) - P_\alpha(p)| \le \varepsilon P_\alpha(p)$. If we further have $P_\alpha(p) \ge 2(n^{1/4}(4\log n)^{-1})^{1-\alpha}$ and $\max_y |p_\varphi(y) - p_\mu(y)| \le 2(\log n)n^{-1/4}$, then,

$$\frac{P_\alpha(p)}{2} \overset{(a)}{\le} \hat{P}_\alpha(x^n) \overset{(b)}{\le} P_\alpha(p_\mu) \overset{(c)}{\le} 2^{1+\alpha} P_\alpha(p_\varphi),$$

where (a) follows from the above assumptions; (b) follows from $A^{\underline{B}} \le A^B$ for any $A, B \ge 0$; and (c) follows from the reasoning below.

- Let S denote the the collection of symbols x such that $p_\mu(x) \le 4(\log n)n^{-1/4}$. Then a convexity argument yields $\sum_{x\in S} \big(p_\mu(x)\big)^\alpha \le (n^{1/4}(4\log n)^{-1})^{1-\alpha}$.
- Using (a), (b), and $P_\alpha(p) \ge 4(n^{1/4}(4\log n)^{-1})^{1-\alpha}$, we immediately obtain $P_\alpha(p_\mu) \ge 2(n^{1/4}(4\log n)^{-1})^{1-\alpha}$ and thus $2\sum_{x\in S} \big(p_\mu(x)\big)^\alpha \le P_\alpha(p_\mu) \le 2\sum_{x\notin S} \big(p_\mu(x)\big)^\alpha$.
- For any symbol $x \notin S$, we have $p_\mu(x) > 4(\log n)n^{-1/4}$. This together with the assumption that $\max_x |p_\varphi(x) - p_\mu(x)| \le 2(\log n)n^{-1/4}$ implies $p_\mu(x) \le 2p_\varphi(x)$.
- Therefore, the inequality $\sum_{x\notin S} \big(p_\mu(x)\big)^\alpha \le 2^\alpha \sum_{x\notin S}(p_\varphi(x))^\alpha \le 2^\alpha P_\alpha(p_\varphi)$ holds.
- Consequently, we establish $P_\alpha(p_\mu(x)) \le 2\sum_{x\notin S} \big(p_\mu(x)\big)^\alpha \le 2^{1+\alpha} P_\alpha(p_\varphi)$.

By the inequality $P_\alpha(p)/2 \le 2^{1+\alpha} P_\alpha(p_\varphi)$ and Corollary 2, if $|\hat{P}^\star_\alpha(x^n) - P_\alpha(p_\varphi)| \ge \varepsilon P_\alpha(p_\varphi)$,

$$p_\varphi(\varphi) \le 2\exp\left(-\Omega_\alpha(\varepsilon^2 n(P_\alpha(p_\varphi))^{1/\alpha})\right) \le 2\exp\left(-\Omega_\alpha(\varepsilon^2 n(P_\alpha(p))^{1/\alpha})\right).$$

Let δ_p denote the quantity on the right-hand side. If we further have $p(\varphi) > \delta_p$, then by definition, $p_\varphi(\varphi) \ge p(\varphi) > \delta_p$. Hence for any x^n with profile φ satisfying both $p(\varphi) > \delta_p$

and $|\hat{P}^{\star}_{\alpha}(x^n) - P_{\alpha}(p)| \le \varepsilon P_{\alpha}(p)$, we must have $|\hat{P}^{\star}_{\alpha}(x^n) - P_{\alpha}(p_{\varphi})| \le \varepsilon P_{\alpha}(p_{\varphi})$. Simple algebra combines the last two inequalities and yields

$$|P_{\alpha}(p) - P_{\alpha}(p_{\varphi})| \le 4\varepsilon P_{\alpha}(p).$$

On the other hand, for a sample $X^n \sim p$ with profile φ', the probability that we have both $p(\varphi') \le \delta_p$ and $|\hat{P}^{\star}_{\alpha}(X^n) - P_{\alpha}(p)| \le \varepsilon P_{\alpha}(p)$ is at most δ_p times the cardinality of the set $\Phi^n_{\alpha,\varepsilon}(p) := \{\varphi(x^n) : x^n \in \mathcal{X}^n \text{ and } |\hat{P}^{\star}_{\alpha}(x^n) - P_{\alpha}(p)| \le \varepsilon P_{\alpha}(p)\}$. Below we complete this argument by finding a tight upper bound on $|\Phi^n_{\alpha,\varepsilon}(p)|$ in terms of its parameters.

For any sequence x^n such that $\varphi(x^n) \in \Phi^n_{\alpha,\varepsilon}(p)$, let $N_{\varphi}(x^n)$ denote the number of prevalences $\varphi_j(x^n)$ that are non-zero. Then by definition, we obtain

$$\sum_{j=0}^{N_{\varphi}(x^n)} \frac{j^{\underline{\alpha}}}{n^{\underline{\alpha}}} \le \sum_{j} \frac{j^{\underline{\alpha}}}{n^{\underline{\alpha}}} \cdot \varphi_j(x^n) = \hat{P}^{\star}_{\alpha}(x^n) \le \frac{3}{2} P_{\alpha}(p).$$

Using the standard falling-factorial identity $((j+1)^{\underline{1+\alpha}} - j^{\underline{1+\alpha}})/(1+\alpha) = j^{\underline{\alpha}}$, we can further simplify the expression on the left-hand side:

$$\sum_{j=0}^{N_{\varphi}(x^n)} \frac{j^{\underline{\alpha}}}{n^{\underline{\alpha}}} = \frac{(N_{\varphi}(x^n)+1)^{\underline{1+\alpha}}}{(1+\alpha)n^{\underline{\alpha}}}.$$

This together with the inequality above yields $N_{\varphi}(x^n) \le T^n_{\alpha}(p) := (3(1+\alpha)n^{\alpha} \cdot P_{\alpha}(p)/2)^{1/(1+\alpha)}$. Further note that each prevalence in $\varphi(x^n) = (\varphi_1(x^n), \ldots, \varphi_n(x^n))$ can only take values in $\lceil n \rfloor := \{0, 1, \ldots, n\}$. Therefore, $|\Phi^n_{\alpha,\varepsilon}(p)|$ is at most the number of $T^n_{\alpha}(p)$-sparse vectors over $\lceil n \rfloor^n$, which admits the following upper bound

$$\binom{n}{T^n_{\alpha}(p)} |\lceil n \rfloor|^{T^n_{\alpha}(p)} \le (n+1)^{2T^n_{\alpha}(p)}.$$

Therefore, for $\delta_p \cdot |\Phi^n_{\alpha,\varepsilon}(p)|$ to be small, it suffices to have

$$\Omega_\alpha(\varepsilon^2 n(P_\alpha(p))^{1/\alpha}) \gg 2T^n_\alpha(p)\log(n+1) = 2(3(1+\alpha)n^\alpha \cdot P_\alpha(p)/2)^{1/(1+\alpha)}\log(n+1),$$

which in turn simplifies to

$$\varepsilon^2 n^{1/(1+\alpha)}(P_\alpha(p))^{1/(\alpha(1+\alpha))} \gg \Theta_\alpha(\log n).$$

Following this and $P_\alpha(p) \geq 4(n^{1/4}(4\log n)^{-1})^{1-\alpha}$, we obtain the following lower bound on n.

$$n \gg \Theta_\alpha((\varepsilon^2|\log\varepsilon|)^{-(1+\alpha)}(P_\alpha(p))^{-1/\alpha}).$$

In this case, the probability bound $\delta_p \cdot |\Phi^n_{\alpha,\varepsilon}(p)|$ is no larger than $1/6$.

Finally, let C denote the collection of sequences x^n with profile φ that do not satisfy $|\hat{P}^\star_\alpha(x^n) - P_\alpha(p)| \leq \varepsilon P_\alpha(p)$ or $\max_x |p_\varphi(x) - \mu_x(x^n)/n| \leq 2(\log n)n^{-1/4}$. By Corollary 2, Lemma 8, and the union bound,

$$\Pr_{X^n\sim p}(X^n \in C) \leq 2\exp\left(-\Omega_\alpha(\varepsilon^2 n(P_\alpha(p))^{1/\alpha})\right) + O\left(\frac{1}{n}\right).$$

For n satisfying the lower-bound inequality above, the right-hand side is again no larger than $1/6$. This completes the proof of the theorem.

2.D Sorted Distribution Estimation

2.D.1 Sorted ℓ_1 Distance and Wasserstein Duality

For convenience, we first restate the theorem.

Theorem 5. *If* $n = \Omega(n(\varepsilon)) = \Omega\left(k/(\varepsilon^2 \log k)\right)$ *and* $\varepsilon \geq n^{-c}$,

$$\Pr(\ell_1^{\text{i}}(p_\varphi, p) \geq \varepsilon) \leq \exp(-\Omega(n^{1/11})).$$

In this section, we relate the estimation of sorted distributions to that of distribution properties through a dual definition of the 1-Wasserstein distance.

Recall that we let $\{p\}$ denote the multiset of probability values of a distribution $p \in \Delta_{\mathcal{X}}$. The sorted ℓ_1 distance between two distributions $p, q \in \Delta_{\mathcal{X}}$ is

$$\ell_1^{\text{i}}(p, q) := \min_{q' \in \Delta_{\mathcal{X}} : \{q'\} = \{q\}} \left\| p - q' \right\|_1,$$

which is invariant under domain-symbol permutations on either p or q.

For two distributions ω, ν over the unit interval $[0,1]$, let $\Gamma'_{\omega,\nu}$ be the collection of distributions over $[0,1] \times [0,1]$ with marginals ω and ν on the first and second factors respectively. The 1-*Wasserstein distance*, also known as the *earth-mover distance*, between ω and ν is

$$\mathcal{W}_1(\omega, \nu) := \inf_{\gamma \in \Gamma'_{\omega,\nu}} \mathop{\mathbb{E}}_{(X,Y) \sim \gamma} |X - Y|.$$

Equivalently, let $\mathcal{L}_1$ denote the collection of real functions that are 1-Lipschitz on $[0,1]$. Through duality, one can also define the 1-Wasserstein distance [KR58] as

$$\mathcal{W}_1(\omega, \nu) = \sup_{f \in \mathcal{L}_1} \left(\mathop{\mathbb{E}}_{X \sim \omega} f(X) - \mathop{\mathbb{E}}_{Y \sim \nu} f(Y) \right).$$

For any $p \in \Delta_{\mathcal{X}}$, let $u_{\{p\}}$ denote the distribution induced by the uniform measure on $\{p\}$. For any distributions $p, q \in \Delta_{\mathcal{X}}$, one can verify [VV16, GR16, HJW18] that

$$\ell_1^{\text{i}}(p, q) = k \cdot \mathcal{W}_1(u_{\{p\}}, u_{\{q\}}) \leq R(p, q).$$

Combining this with the dual definition of $\mathcal{W}_1$, we obtain

$$\ell_1^{\mathrm{i}}(p,q) = k \cdot \sup_{f \in \mathcal{L}_1} \left(\mathop{\mathbb{E}}_{X \sim u_{\{p\}}} f(X) - \mathop{\mathbb{E}}_{Y \sim u_{\{q\}}} f(Y) \right) = \sup_{f \in \mathcal{L}_1} \left(\sum_x f(p(x)) - \sum_x f(q(x)) \right).$$

2.D.2 Proof of Theorem 5

For a real function $f \in \mathcal{L}_1$, we denote by $f(p) := \sum_x f(p(x))$ the corresponding additive symmetric property. The previous reasoning also shows that for any $p, q \in \Delta_{\mathcal{X}}$,

$$R(p,q) \geq \ell_1^{\mathrm{i}}(p,q) \geq |f(p) - f(q)|.$$

Therefore, property f is 1-Lipschitz on $(\Delta_{\mathcal{X}}, R)$.

Set $n := \sup_{f \in \mathcal{L}_1} n_f(\varepsilon)$. The results in [HJW18] imply that if $\varepsilon > n^{-0.3}$,

$$n = \Theta\left(\frac{k}{\varepsilon^2 \log k} \right).$$

Clearly, we only need to consider $\varepsilon \leq 2$, implying $k = O(n \log n)$. Let α, γ be absolute constants in $[1/100, 1/6)$ and $\varepsilon > 0$ be an error parameter.

By the proof of Theorem 1 in Section 2.B.2, for any distribution $p \in \Delta_{\mathcal{X}}$ and $X^{n/\alpha} \sim p$, with probability at least $1 - 2\exp\left(-4n^{1-2\gamma}\right)$, the PML (or APML) plug-in estimator will satisfy

$$|f(p) - f(p_{\varphi(X^{n/\alpha})})| < \varepsilon\,(2 + o(1)) + O(n^{-c_1/2} \log^2 n) + 4n^{1-\gamma}\tau(n),$$

where $c_1 \in (1/100, 1/32]$, $c_2 = 1/2 + 6c_1$, and $\tau(n) = O\left(n^{\alpha c_2 + (2-\alpha)c_1 - 1} \log^3 n \right)$. Additionally, in the previous section, we have proved that

$$\ell_1^{\mathrm{i}}(p,q) = \sup_{f \in \mathcal{L}_1} (f(p) - f(q)) = \sup_{f \in \mathcal{L}_1} |f(p) - f(q)|.$$

Though it seems that the above inequality and equation imply the optimality of PML (since f is chosen arbitrarily), such direct implication actually does not hold. The reason is a little bit subtle: The inequality on $|f(p) - f(p_{\varphi(X^{n/\alpha})})|$ holds for any fixed function f and $p \in \Delta_{\mathcal{X}}$, while the function that achieves the corresponding supremum in

$$\sup_{f \in \mathcal{L}_1} \left| f(p) - f(p_{\varphi(X^{n/\alpha})}) \right| = \ell_1^{\mathrm{i}} \left(p, p_{\varphi(X^{n/\alpha})} \right)$$

depends on both p and $X^{n/\alpha}$, and hence is a random function. To address this discrepancy, we provide a more involved argument below.

Let f be a function in $\mathcal{L}_1$. Without loss of generality, we also assume that $f(0) = 0$. Let $\eta \in (0,1)$ be a threshold parameter to be determined later. An η-*truncation* of f is a function

$$f_\eta(z) := f(z)\mathbb{1}_{z \le \eta} + f(\eta)\mathbb{1}_{z > \eta}.$$

One can easily verify that $f_\eta \in \mathcal{L}_1$. Next, we find a finite subset of $\mathcal{L}_1$ so that the η-truncation of any $f \in \mathcal{L}_1$ is close to at least one of the functions in this subset.

For a parameter $s > 3$ to be chosen later. Partition the interval $[0, \eta]$ into s disjoint sub-intervals of equal length, and define the sequence of end points as $z_j := \eta \cdot j/s, j \in \lceil s \rfloor$ where $\lceil s \rfloor := \{0, 1, \ldots, s\}$. Then, for each $j \in \lceil s \rfloor$, we find the integer j' such that $|f_\eta(z_j) - z_{j'}|$ is minimized and denote it by j^*. Since f_η is 1-Lipschitz, we must have $|j^*| \in \lceil j \rfloor$. Finally, we connect the points $Z_j := (z_j, z_{j^*})$ sequentially. This curve is continuous and corresponds to a particular η-truncation $\tilde{f}_\eta \in \mathcal{L}_1$, which we refer to as the *discretized* η-*truncation* of f. Intuitively, we have constructed an $(s+1) \times (s+1)$ grid and "discretized" function f by finding its closest approximation in $\mathcal{L}_1$ whose curve only consists of edges and diagonals of the grid cells. By construction,

$$\max_{z \in [0,1]} |f_\eta(z) - \tilde{f}_\eta(z)| \le \eta/s.$$

Therefore, for any $p \in \Delta_{\mathcal{X}}$, the corresponding properties of f_η and $\tilde{f}_\eta$ satisfy

$$|f_\eta(p) - \tilde{f}_\eta(p)| \le k \cdot \eta / s.$$

Note that $|j^*| \in \lceil j \rfloor$ for all $j \in \lceil s \rfloor$, and $\tilde{f}_\eta(z) = z_{s^*}$ for $z \ge \eta$. While there are infinitely many η-truncations, the cardinality of the discretized η-truncations of functions in $\mathcal{L}_1$ is at most

$$\prod_{j=0}^{s}(2j+1) = (s+1)\prod_{j=0}^{s-1}(2j+1)(2s-2j+1) \le (s+1)^{2s+1} = e^{(2s+1)\log(s+1)} \le e^{3s\log s}.$$

Consider any $p \in \Delta_{\mathcal{X}}$ and $X^{n/\alpha} \sim p$ with a profile φ. Consolidate the previous results, and apply the union bound and triangle inequality. With probability at least $1 - 2\exp\left(3s\log s - 4n^{1-2\gamma}\right)$, the PML plug-in estimator will satisfy

$$\begin{aligned} |f_\eta(p) - f_\eta(p_\varphi)| &\le |f_\eta(p) - \tilde{f}_\eta(p)| + |\tilde{f}_\eta(p) - \tilde{f}_\eta(p_\varphi)| + |\tilde{f}_\eta(p_\varphi) - f_\eta(p_\varphi)| \\ &\le 2k \cdot \eta / s + \varepsilon\,(2 + o(1)) + O(n^{-c_1/2}\log^2 n) + 4n^{1-\gamma}\tau(n), \end{aligned}$$

for *all* functions f in $\mathcal{L}_1$.

Next we consider the "second part" of a function $f \in \mathcal{L}_1$, namely,

$$\bar{f}_\eta(z) := f(z) - f_\eta(z) = (f(z) - f(\eta))\mathbb{1}_{z>\eta}.$$

Again, we can verify that $\bar{f}_\gamma \in \mathcal{L}_1$. To establish the corresponding guarantees, we make use of the following result. Since the profile probability $p(\varphi)$ is invariant to symbol permutation, for our purpose, we can assume that $p(y) \le p(z)$ iff $p_\varphi(x) \le p_\varphi(y)$, for all $x, y \in \mathcal{X}$. Under this assumption, the next lemma, which follows from the consistency results in [OSVZ11, AGZ17], relates p_φ to p. Let $\gamma' \in (0, 1/4)$ be an absolute constant to be determined later. Then,

Lemma 9. *For any distribution p and sample $X^m \sim p$ with profile φ,*

$$\Pr\left(\max_x |p_\varphi(x) - p(x)| > m^{\gamma'-1/4}\right) = O\left(m^{1/4}\exp(-\Omega(m^{1/2+2\gamma'}))\right).$$

Simply following the proofs in [OSVZ11, AGZ17], we obtain: Changing $1/4$ to any (fixed) number greater than $1/6$, the above lemma also holds for APML with $m^{1/2+2\gamma'}$ replaced by $m^{2/3+2\gamma'}$.

For $m = n/\alpha$, with probability at least $1 - O\left((n/\alpha)^{1/4}\exp(-\Omega((n/\alpha)^{1/2+2\gamma'}))\right)$,

$$\begin{aligned}
|\bar{f}_\eta(p) - \bar{f}_\eta(p_\varphi)| &= |\sum_x \bar{f}_\eta(p(x)) - \bar{f}_\eta(p_\varphi(x))| \\
&\le \sum_{x: p(x)>\eta \text{ or } p_\varphi(x)>\eta} |\bar{f}_\eta(p(x)) - \bar{f}_\eta(p_\varphi(x))| \\
&\le \sum_{x: p(x)>\eta \text{ or } p_\varphi(x)>\eta} |p(x) - p_\varphi(x)| \\
&\le (2/\eta)(n/\alpha)^{\gamma'-1/4},
\end{aligned}$$

for *all* functions f in $\mathcal{L}_1$.

Consolidate the previous results. By the triangle inequality and the union bound, with probability at least $1 - 2\exp\left(3s\log s - 4n^{1-2\gamma}\right) - O\left((n/\alpha)^{1/4}\exp(-\Omega((n/\alpha)^{1/2+2\gamma'}))\right)$,

$$\begin{aligned}
|f(p) - f(p_\varphi)| &\le |f_\eta(p) - f_\eta(p_\varphi)| + |\bar{f}_\eta(p) - \bar{f}_\eta(p_\varphi)| \\
&\le 2k\eta/s + \varepsilon(2 + o(1)) + O(n^{-c_1/2}\log^2 n) + 4n^{1-\gamma}\tau(n) + (2/\eta)(n/\alpha)^{\gamma'-1/4},
\end{aligned}$$

for *all* functions f in $\mathcal{L}_1$. Now we can conclude that $\ell_1^i\left(p, p_\varphi\right)$ is also at most the error bound on the right-hand side. The reason is straightforward: Since with high probability, the above guarantee holds for all functions in $\mathcal{L}_1$, it must also hold for the function that achieves the

supremum in

$$\sup_{f\in\mathcal{L}_1}\left|f(p)-f(p_\varphi)\right|=\ell_1^{\mathrm{i}}\left(p,p_\varphi\right).$$

It remains to make sure that all the quantities in the error bound except $\varepsilon(2+o(1))$ vanish with n, and the probability bound converges to 1 as n increases. Recall that $k=O(n\log n)$, $c_1\in(1/100,1/25]$, $c_2=1/2+6c_1$, and $\tau(n)=O\left(n^{\alpha c_2+(2-\alpha)c_1-1}\log^3 n\right)$.

By direct computation, we can choose $\alpha=1/100$, $c_1=1/26$, $\gamma'=1/200$, $\gamma=(5/2+5\alpha)c_1+\alpha/2$, $s=n^{\gamma'+3/4+c_1}$, and $\eta=n^{\gamma'-1/4+c_1/2}$. Note that this is just one possible set of parameters. Given this choice, we have

$$\ell_1^{\mathrm{i}}\left(p,p_\varphi\right)\le\varepsilon(2+o(1))+O(n^{-c_1/2}\log^3 n),$$

with probability at least $1-\exp(-\Omega(n^{1/2}))$. Additionally, the equation

$$\sup_{f\in\mathcal{L}_1}\left|f(p)-f(p_\varphi)\right|=\ell_1^{\mathrm{i}}\left(p,p_\varphi\right)$$

clearly yields that $n(\varepsilon)\ge\sup_{f\in\mathcal{L}_1}n_f(\varepsilon)$. Hence for $\varepsilon\ge O(n^{-c_1/2}\log^4 n)$,

$$n(p_\varphi,(2+o(1))\varepsilon)\le 100n(\varepsilon).$$

2.E Uniformity Testing

2.E.1 PML-Based Tester

Let ε be an arbitrary accuracy parameter and $\mathcal{X}$ be a finite set. Let p_u denote the uniform distribution over $\mathcal{X}$. Given sample access to an unknown distribution $p\in\Delta_{\mathcal{X}}$, the uniformity

testing distinguishes between the null hypothesis

$$H_0 : p = p_u$$

and the alternative hypothesis

$$H_1 : \|p - p_u\|_1 \geq \varepsilon.$$

After a sequence of research works [GR00, BFF+01, Pan08, AJOS13a, CDVV14, VV17, DKN15, ADK15, DK16, DGPP18], it is shown that to achieve a $k^{-\Theta(1)}$ bound on the error probability, this task requires a worst-case sample size of order $\sqrt{k \log k}/\varepsilon^2$. The uniformity tester $T_{\text{PML}}(X^n)$ in Figure 2.8 is purely based on PML, and takes as input parameters k and ε, and a sample $X^n \sim p$.

Input: parameters $k, \varepsilon,$ and a sample $X^n \sim p$ with profile φ.
1. If $\max_x \mu_x(X^n) \geq 3 \max\{1, n/k\} \log k$, return 1;
2. Elif $\left\|p_\varphi - p_u\right\|_2 \geq 3\varepsilon/(4\sqrt{k})$, return 1;
3. Else return 0.

Figure 2.8: Uniformity tester T_{PML}

In the rest of this section, we establish the following theorem.

Theorem 6. *If $\varepsilon = \tilde{\Omega}(k^{-1/4})$ and $n = \tilde{\Omega}(\sqrt{k}/\varepsilon^2)$, the tester $T_{PML}(X^n)$ will be correct with probability at least $1 - k^{-2}$. The tester also distinguishes between $p = p_u$ and $\|p - p_u\|_2 \geq \varepsilon/\sqrt{k}$.*

2.E.2 Proof of Theorem 6

Assume that $\varepsilon \geq (\log k)/k^{1/4}$. For a sample $X^n \sim p_u$, the multiplicity of each symbol x follows a binomial distribution $\text{bin}(n, k^{-1})$ with mean n/k. The following lemma [Che81] bounds the tail probability of a binomial random variable.

Lemma 10. *For a binomial random variable Y with mean M and any $t \geq 1$,*

$$\Pr(Y \geq (1+t)M) \leq \exp(-t(2/t+2/3)^{-1}M).$$

Applying the above lemma to $Y = \mu_x(X^n)$ and $t = 3\max\{k/n, 1\}\log k$ immediately yields that $\Pr(\mu_x(X^n) \geq (1+t)n/k) \leq k^{-3}$. By symmetry and the union bound, we then have $\Pr\left(\max_x \mu_x(X^n) \geq (1+t)n/k\right) \leq k^{-2}$. In the subsequent discussion, we denote by $\Phi^n_{\mathcal{X}}$ the profile set $\{\varphi(x^n) : x^n \in \mathcal{X}^n \text{ and } \max_x \mu_x(x^n) < (1+t)n/k\}$.

Consider the problem of estimating the ℓ_2-distance between an unknown distribution and the uniform distribution p_u, for which we have the following result [Gol17].

Lemma 11. *There is a profile-based estimator $\hat{\ell}_2$ such that for any $\varepsilon_0 \leq k^{-1/2}$, $n = \Omega(k^{-1/2}/\varepsilon_0^2)$, $p \in \Delta_{\mathcal{X}}$ satisfying $P_2(p) = O(k^{-1})$, and $X^n \sim p$,*

- *if $\|p - p_u\|_2 > \varepsilon_0$, then $\hat{\ell}_2(X^n) \geq 0.9\varepsilon_0$,*
- *if $\|p - p_u\|_2 < \varepsilon_0/2$, then $\hat{\ell}_2(X^n) \leq 0.6\varepsilon_0$,*

with probability at least $2/3$.

Set $\varepsilon_0 = \varepsilon/\sqrt{k}$ in the above lemma. Then, by the sufficiency of profiles and the standard median trick, there exists another profile-based estimator $\hat{\ell}_2^\star$ that under the same conditions, provides the estimation guarantees stated above, with probability at least $1-\delta$ for $\delta := 2\exp(-\Omega(n\varepsilon^2/\sqrt{k}))$. Scaling ε_0 by positive absolute constant factors yields: If $\|p - p_u\|_2 > 0.67\varepsilon_0$, then $\hat{\ell}_2(X^n) \leq 0.6\varepsilon_0$ with probability at most δ; if $\|p - p_u\|_2 < 0.75\varepsilon_0$, then $\hat{\ell}_2(X^n) \geq 0.9\varepsilon_0$ with probability at most δ.

Let φ' be a profile. If we further have $p(\varphi') > \delta$, then by definition, $p_{\varphi'}(\varphi') \geq p(\varphi') > \delta$. Hence for any x^n with profile φ', if $\|p - p_u\|_2 > \varepsilon_0$, we must have both $\hat{\ell}_2(x^n) \geq 0.9\varepsilon_0$ and $\left\|p_{\varphi'} - p_u\right\|_2 \geq 0.75\varepsilon_0$; if $\|p - p_u\|_2 < \varepsilon_0/2$, we must have both $\hat{\ell}_2(x^n) \leq 0.6\varepsilon_0$ and $\left\|p_{\varphi'} - p_u\right\|_2 \leq 0.67\varepsilon_0$.

On the other hand, for a sample $X^n \sim p$ with profile φ, the probability that we have both $p(\varphi) \le \delta$ and $\varphi \in \Phi^n_{\mathcal{X}}$ is at most δ times the cardinality of the set $\Phi^n_{\mathcal{X}}$. By definition, if $\varphi \in \Phi^n_{\mathcal{X}}$, then $\varphi_i = 0$ for $i \ge (1+t)n/k$. In addition, each φ_i can only take values in $\lceil k \rfloor = \{0,1,\ldots,k\}$, implying that $|\Phi^n_{\mathcal{X}}| \le |\lceil k \rfloor|^{(1+t)n/k} \le \exp(6\max\{n/k,1\}\log^2 k)$. Therefore, we obtain the following upper bound on the probability of interest: $\delta \cdot |\Phi^n_{\mathcal{X}}| \le 2\exp(-\Omega(n\varepsilon^2/\sqrt{k}) + 6\max\{n/k,1\}\log^2 k)$. In order to make the probability bound vanish, we need to consider two cases: $n \le k$ and $n > k$. If $n \le k$, it suffices to have $n \gg (\log^2 k)\sqrt{k}/\varepsilon^2$; If $n > k$, it suffices to have $\varepsilon \gg (\log k)/k^{1/4}$. In both cases, the probability bound is at most $\exp(-\log^2 k)$.

Next, consider estimating the power sum $P_2(p)$, which is at least $k^{-1/2}$ for $p \in \Delta_{\mathcal{X}}$. By Corollary 2, there is a profile-based estimator $\hat{P}_2^\star$ such that $\Pr_{X^n \sim p}(|\hat{P}_2^\star(X^n) - P_2(p)| \ge (\varepsilon/8) \cdot P_2(p)) \le 2\exp(-\Omega(n\varepsilon^2/\sqrt{k})) = \delta$. Following the same derivations as above and in Section 2.C.3 with $\Phi^n_{\alpha,\varepsilon}(p)$ replaced by $\Phi^n_{\mathcal{X}}$, we establish that

$$\Pr\left(|P_2(p_\varphi) - P_2(p)| > P_2(p)/2 \text{ and } \varphi \in \Phi^n_{\mathcal{X}}\right) \le \delta \cdot |\Phi^n_{\mathcal{X}}| \le \exp(-\log^2 k).$$

Now we are ready to characterize the performance of the tester $T_{\text{PML}}(X^n)$. For clarity, we divide our analysis into two parts based on which hypothesis is true.

- **Case 1:** The null hypothesis H_0 is true, i.e., $p = p_u$.
  - **Step 1:** By Lemma 10 and its implications, given $p = p_u$, the probability of failure at this step is at most $\Pr_{X^n \sim p_u}(\exists x \in \mathcal{X} \text{ s.t. } \mu_x(X^n) \ge (1+t)n/k) \le k^{-2}$.
  - **Step 2:** Note that $P_2(p) = k^{-1}$ and $\|p - p_u\|_2 = 0$, and recall that $\varphi = \varphi(X^n)$. The tester accepts H_1 in this step iff $\varphi \in \Phi^n_{\mathcal{X}}$ and $\left\|p_\varphi - p_u\right\|_2 \ge 0.75\varepsilon_0$. By Lemma 11 and the subsequent arguments, this happens with probability at most $\exp(-\log^2 k)$.
  - **Step 3:** The tester always accepts H_0 in this step. Hence by the union bound, if the null hypothesis H_0 is true, then the tester succeeds with probability at least $1 - k^{-2}$.

- **Case 2:** The alternative hypothesis H_1 is true, i.e., $\|p - p_u\|_1 \geq \varepsilon$.

  - **Step 1 to 2:** The tester accepts H_1 if the conditions in either Step 1 or Step 2 are satisfied, and hence incurs no error.

  - **Step 3:** By the value of $P_2(p)$, we further divide our analysis into two parts:

    * If $P_2(p) \geq 10k^{-1}$, then $\left\|p_\varphi - p_u\right\|_2 < 0.75\varepsilon/\sqrt{k}$ implies that $P_2(p_\varphi) < 1.6k^{-1}$ and $|P_2(p_\varphi) - P_2(p)| > P_2(p)/2$. Hence, the tester accepts H_0 only if both $|P_2(p_\varphi) - P_2(p)| > P_2(p)/2$ and $\varphi \in \Phi^n_{\mathcal{X}}$ happen, whose probability, by the above disscusion, is at most $\exp(-\log^2 k)$.

    * If $P_2(p) < 10k^{-1}$, then all the conditions in Lemma 11 are satisfied. In addition, by the Cauchy-Schwarz inequality, we have $\|p - p_u\|_2 \geq \|p - p_u\|_1 \cdot k^{-1/2} \geq \varepsilon \cdot k^{-1/2}$. The tester accepts H_0 iff both $\left\|p_\varphi - p_u\right\|_2 < 0.75\varepsilon \cdot k^{-1/2}$ and $\varphi \in \Phi^n_{\mathcal{X}}$ hold, which happen, by Lemma 11 and the subsequent arguments, with probability at most $\exp(-\log^2 k)$.

This completes the proof of the theorem.

2.F Proof of Lemma 3

The proof closely follows that of Proposition 6.19 in [Val12] (page 131–136), which we refer to as *the proposition's proof*. Note that in the work [Val12], the definitions of k and n are swapped, i.e., k stands for the sample size, and n denotes the alphabet size. For consistency, we still keep our notation.

Recall that we set $t_n := 2n^{-c_1} \log n$ and $\alpha \in (0,1)$, and define

$$\beta_i := (1 - e^{-t_n \alpha i}) f\left(\frac{(i+1)\alpha}{n}\right) \frac{n}{(i+1)\alpha} + \sum_{\ell=0}^{i} z_\ell (1-t_n)^\ell \alpha^\ell (1-\alpha)^{i-\ell} \binom{i}{\ell}.$$

for any $i \le n$, and $\beta_i := \beta_n$ for $i > n$. Let $w(i)$ denote the first quantity on the right-hand side, and $w := (w(0), w(1), \ldots)$ be the corresponding vector. Similarly, let $\tilde{z}_\alpha(i)$ denote the second quantity on the right-hand side, and $\tilde{z}_\alpha$ be the corresponding vector. Assume that $v \le \log^2 n$.

First part of the proposition's proof remains unchanged, which corresponds to the content from page 131 to the second last paragraph on page 132, showing that

$$\sqrt{\alpha}\,\|\tilde{z}_\alpha\|_2 = O(n^{\alpha c_2 + (1-\alpha)c_1} \cdot \log^3 n).$$

The assumption that $\alpha \in [1/100, 1)$ implies $\sqrt{\alpha} \ge 1/10$, and hence we have $|\tilde{z}_\alpha(i)| \le \|\tilde{z}_\alpha\|_2 = O(n^{\alpha c_2 + (1-\alpha)c_1} \cdot \log^3 n)$. Recall that for lemma 2 to hold, the coefficients β_i must satisfy the following two conditions,

1. $|\varepsilon(y)| \le a' + b'/y$,
2. $|\beta^\star_j - \beta^\star_\ell| \le c'\sqrt{j/n}$ for any j and ℓ such that $|j - \ell| \le \sqrt{j}\log n$,

where $\varepsilon(y) := f(y)/y - e^{-ny}\sum_{i \ge 0} \beta_i \cdot (ny)^i / i!$, and $\beta^\star_i := \beta_{i-1} \cdot i/n, \forall i \ge 1$, and $\beta^\star_0 := 0$.

We first consider the second condition and find a proper parameter c'.

Our objective is to find $c' > 0$ such that $c' > \sqrt{n/j}\,|\beta^\star_j - \beta^\star_\ell|$. By the triangle inequality,

$$\sqrt{\frac{n}{j}}|\beta^\star_j - \beta^\star_\ell| \le \sqrt{\frac{n}{j}}\left|\frac{j}{n}\tilde{z}_\alpha(j-1) - \frac{\ell}{n}\tilde{z}_\alpha(\ell-1)\right| + \sqrt{\frac{n}{j}}\left|\frac{j}{n}w(j-1) - \frac{\ell}{n}w(\ell-1)\right|.$$

We bound the two quantities on the right-hand side separately and consider two cases for each. If both j and ℓ are at most $400n^{c_1}$, then

$$\sqrt{\frac{n}{j}}\left|\frac{j}{n}\tilde{z}_\alpha(j-1) - \frac{\ell}{n}\tilde{z}_\alpha(\ell-1)\right| \le O(n^{c_1/2 - 1/2}) \cdot \max_i |z_\alpha(i)| \le O(n^{\alpha c_2 + (3/2-\alpha)c_1 - 1/2}\log^3 n).$$

Recall that $|z_\ell| \le v \cdot n^{c_2}, \forall \ell \ge 0$. If one of j and ℓ is larger than $400n^{c_1}$, say $j > 400n^{c_1}$, then

$$\begin{aligned}
\sqrt{\frac{n}{j}}\left|\frac{j}{n}\tilde{z}_\alpha(j-1)\right| &\le \sqrt{\frac{j}{n}}\sum_{\ell=0}^{j-1}|z_\ell|(1-t_n)^\ell\alpha^\ell(1-\alpha)^{j-1-\ell}\binom{j-1}{\ell} \\
&\le \sqrt{j}n^{c_2-1/2}(\log^2 n)\sum_{\ell=0}^{j-1}(1-t_n)^\ell\alpha^\ell(1-\alpha)^{j-1-\ell}\binom{j-1}{\ell} \\
&= \sqrt{j}n^{c_2-1/2}(\log^2 n)(1-t_n\alpha)^{j-1} \\
&\le \sqrt{j}n^{c_2-1/2}(\log^2 n)(1-\log n/(50n^{c_1}))^{400n^{c_1}} \\
&\le \sqrt{j}n^{c_2-1/2}(\log^2 n)n^{-8}.
\end{aligned}$$

For $j < 2n^2$, the last quantity is at most n^{-1}. For $j > 2n^2$, we have $\ell > n^2$ and hence

$$\sqrt{\frac{n}{j}}\left|\frac{j}{n}\tilde{z}_\alpha(j-1) - \frac{\ell}{n}\tilde{z}_\alpha(\ell-1)\right| = \sqrt{\frac{n}{j}}\,|j-\ell|\,\tilde{z}_\alpha(n-1) \le \sqrt{n}(\log n)n^{-1} = (\log n)n^{-1/2}.$$

Similarly, we can bound the other quantity, i.e.,

$$\sqrt{\frac{n}{j}}\left|\frac{j}{n}w(j-1) - \frac{\ell}{n}w(\ell-1)\right| = \sqrt{\frac{n}{\alpha^2 j}}\left|(1-e^{-t_n\alpha(j-1)})f\left(\frac{j\alpha}{n}\right) - (1-e^{-t_n\alpha(\ell-1)})f\left(\frac{\ell\alpha}{n}\right)\right|.$$

Since f (the property) is 1-Lipschitz on $(\Delta_{\mathcal{X}}, R)$ and $f(p) = 0$ if $p(x) = 1$ for some $x \in \mathcal{X}$, one can verify that $|f(x)| \le x|\log x| \le e^{-1}$ and $|f(x)/x - f(y)/y| \le |\log(x/y)|$ for $x, y \in [0,1]$ (the corresponding real function). We consider two cases and bound the quantity of interest. If $j \ge \sqrt{n}$,

$$\sqrt{\frac{n}{\alpha^2 j}}\left|(1-e^{-t_n\alpha(j-1)})f\left(\frac{j\alpha}{n}\right)\right| \le \sqrt{\frac{n}{\alpha^2 j}}\left|f\left(\frac{j\alpha}{n}\right)\right| \le \sqrt{\frac{n}{\alpha^2 j}}\frac{j\alpha}{n}\log\left(\frac{j\alpha}{n}\right) \le O(n^{-1/4}\log n).$$

The same bound also applies to the other term where j is replaced by ℓ. If $j > \sqrt{n}$, then $e^{-t_n\alpha(j-1)} \le \exp(-2\alpha(\log n)n^{1/2-c_1}) = O(n^{-2})$. Analogously, the same upper bound holds for

the other term $e^{-t_n\alpha(\ell-1)}$. Hence, we ignore these two terms and consider only

$$\begin{aligned}
\sqrt{\frac{n}{\alpha^2 j}}\left|f\left(\frac{j\alpha}{n}\right)-f\left(\frac{\ell\alpha}{n}\right)\right| &\le \sqrt{\frac{j}{n}}\left|\frac{n}{j\alpha}f\left(\frac{j\alpha}{n}\right)-\frac{n}{\ell\alpha}f\left(\frac{\ell\alpha}{n}\right)\right|+\sqrt{\frac{j}{n}}\left|\frac{n}{j\alpha}-\frac{n}{\ell\alpha}\right|f\left(\frac{\ell\alpha}{n}\right)\\
&\le \sqrt{\frac{j}{n}}\left|\log\frac{j}{\ell}\right|+\sqrt{\frac{j}{n}}\left|\frac{n}{j\alpha}-\frac{n}{\ell\alpha}\right|f\left(\frac{\ell\alpha}{n}\right)\\
&\le \sqrt{\frac{j}{n}}\frac{|j-\ell|}{j}+\frac{\sqrt{jn}}{\alpha}\frac{|j-\ell|}{j\ell}f\left(\frac{\ell\alpha}{n}\right)\\
&\le \sqrt{\frac{j}{n}}\frac{|j-\ell|}{j}+\sqrt{\frac{j}{n}}\frac{|j-\ell|}{j}\left|\log\left(\frac{\ell\alpha}{n}\right)\right|\\
&\le \frac{\log n}{\sqrt{n}}+\frac{\log n}{n}\left|\log\left(\frac{\ell\alpha}{n}\right)\right|\\
&= O(n^{-1/2}\log n).
\end{aligned}$$

Since $\alpha c_2+(3/2-\alpha)c_1\le 1/4$, we have $O(n^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^3 n)=O(n^{-1/4}\log^3 n)$. Hence, we can set the latter quantity to be c'. The above derivations also show that

$$|w(i)|=\left|(1-e^{-t_n\alpha i})f\left(\frac{(i+1)\alpha}{n}\right)\frac{n}{(i+1)\alpha}\right|\le\left|\log\left(\frac{(i+1)\alpha}{n}\right)\right|=O(\log n).$$

Together with $\beta_i=w(i)+\tilde{z}_\alpha(i)$ and $|\tilde{z}_\alpha(i)|=O(n^{\alpha c_2+(1-\alpha)c_1}\cdot\log^3 n)$, this inequality implies

$$|\beta_i|\le O(n^{\alpha c_2+(1-\alpha)c_1}\log^3 n).$$

It remains to analyze the first condition of Lemma 2 and find proper values for a' and b'. For this part, the corresponding proof in [Val12] also holds for $\alpha\in[1/100,1/2]$ (page 134 to the second last paragraph on page 135), hence no change is needed. One thing to note is that $1/\alpha$ and $1/\sqrt{\alpha}$ are both $O(1)$. For some $a'',b''\ge 0$ such that $a''+b''k\le v$, we can set $a'=a''+O(n^{-c_1/2}\log^2 n)$ and $b'=b''(1+O(n^{-c_1}\log n))$. The proof of Lemma 3 is complete.

Chapter 3

Doubly Competitive Distribution Estimation

3.1 Introduction

Estimating large-alphabet distributions from their samples is a fundamental statistical-learning staple. Over the past few decades, distribution estimation has found numerous applications, ranging from language modeling [CG99] to biological studies [AIS+08], and has been extensively studied. In the following subsections, we formalize the discussion and present major research frameworks used in the field.

3.1.1 Distribution Estimation

Let Δ_k denote the collection of distributions over the discrete *alphabet* $[k] := \{1, \ldots, k\}$. Let $[k]^*$ be the set of finite-length sequences over $[k]$. An *estimator* is a mapping $\hat{p} : [k]^* \to \Delta_k$ that associates with every sequence x^n a distribution $\hat{p}(x^n) \in \Delta_k$. Let $X^n := X_1, \ldots, X_n$ be an i.i.d. sample sequence from an unknown p. Our objective is to find an estimator $\hat{p}$ such that $\hat{p}(X^n)$ approximates p well.

Specifically, for two distributions $p, q \in \Delta_k$, let $\ell(p,q)$ be the *loss* when approximating distribution p by estimate q. The loss of estimating p by $\hat{p}(X^n)$ is therefore $\ell(p, \hat{p}(X^n))$. We also consider the expected loss, known as *risk*,

$$r_n^\ell(p, \hat{p}) := \mathbb{E}_{X^n \sim p} L(p, \hat{p}(X^n)).$$

The two most important losses for distribution estimation are the KL-divergence $D(p \parallel q) := \sum_{x \in [k]} p_x \log \frac{p_x}{q_x}$, and the ℓ_1-distance $|p-q| := \sum_{x \in [k]} |p_x - q_x|$. *We study mainly the KL-loss, hence abbreviate r^{KL} as simply r.*

Next, we formalize the uncertainty about the distribution and the three common measures for the approximation quality: min-max, structural, and competitive estimation.

3.1.2 Previous Works

Min-max While the underlying distribution p is unknown, it often belongs to a known distribution collection $\mathcal{P}$. The *worst-case risk* of an estimator $\hat{p}$ over all distributions in $\mathcal{P}$ is

$$r_n^\ell(\mathcal{P}, \hat{p}) := \max_{p \in \mathcal{P}} r_n^\ell(p, \hat{p}),$$

and the minimal possible worst-case risk for $\mathcal{P}$, incurred by any estimator, is the *min-max risk*,

$$r_n^\ell(\mathcal{P}) := \min_{\hat{p}} r_n^\ell(\mathcal{P}, \hat{p}) = \min_{\hat{p}} \max_{p \in \mathcal{P}} r_n^\ell(p, \hat{p}).$$

The most classical and widely-studied class of distributions is simply the set Δ_k of all discrete distributions. The problem of determining $r_n^\ell(\Delta_k)$ up to the first order was introduced by [Cov72] and studied in a sequence of papers [KT81, BFSS02, Pan05]. Among the many results on the topic, [BS04] showed that for KL-divergence, as $n/k \to \infty$, the min-max KL-risk satisfies $r_n(\Delta_k) = (1+o(1))\frac{k-1}{2n}$, achieved by a variant of the add-3/4 estimator. On the other hand,

[Pan05] proved that as $k/n \to \infty$, the optimal KL-risk becomes $r_n(\Delta_k) = (1+o(1))\log\frac{k}{n}$, which is achieved by add-constant estimators. Similar results for other loss measures like ℓ_1-distance can be found in [KOPS15].

Beyond min-max The success of add-constant estimators in achieving the classical min-max risks does not extend to practical applications. One possible explanation is that practical distributions, like power-law, or Poisson, are often rather simple and can be estimated more efficiently and accurately than the worst distribution targeted by the min-max paradigm. The desire to construct estimators that perform better on practical distributions has led to the following two frameworks.

Structural Instead of considering arbitrary underlying distributions, the structural approach focuses on learning distributions that posses a natural structure, such as monotonicity, log-concavity, and m-modality. In many cases, structural assumptions lead to more effective estimators that provably perform better on the corresponding distribution classes.

For example, [KOPS15] showed that for fixed k, as n increases, the empirical estimator achieves the min-max ℓ_1-risk over Δ_k,

$$r_n^{\ell_1}(\Delta_k) = (1+o(1))\sqrt{\frac{2(k-1)}{\pi n}}.$$

In many practical applications, the alphabet k is often large, hence several papers considered structured distributions [ADLS17, DKS16b, Kam14, CDSS13, Das12, JW09, FOS08]. For example, for the collection $\mathcal{M}_k^{t,m}$ of t-mixture m-modal distributions over $[k]$, more sophisticated estimators, e.g., [ADLS17] attain

$$r_n^{\ell_1}(\mathcal{M}_k^{t,m}) = \Theta\left(\frac{tm\log k}{n}\right)^{1/3},$$

which for $k/\log k \gg n^{1/3}(tm)^{2/3}$, is lower than $r_n^{\ell_1}(\Delta_k)$.

Drawbacks The structural approach leverages the structure assumptions to design more

efficient estimators, thus has the drawback of relying on the hypothetical models.

For example, to learn t-mixture m-modal distributions efficiently as above, one needs to ensure the correctness of the structure assumption and know both t and m up to constant factors. While it may seem possible to use hypothesis testing to find the best parameters, existing work on distribution property testing shows that even testing whether a distribution is m-modal requires a non-trivial number of samples [CDGR18]. Hence, when t and m are relatively large, finding the best parameters may require many samples.

In addition, many structures possessed by real-world distributions, for example, mixtures of log-concave and log-convex, have not been addressed before.

Competitive Instead of relying on often-uncertain structural assumptions, the competitive distribution estimation framework takes a different view and aims to design universally near-optimal estimators. Any reasonable estimator for i.i.d. distributions would assign the same probability to all symbols appearing the same number of times in the sample, and we let Q_{nat} denote this collection of *natural estimators*.

Our objective is to design a distribution estimator $\hat{p}$ that estimates every distribution nearly as well as the best estimator designed with prior knowledge of the true distribution p, but is restricted to be natural. Specifically, for any distribution $p \in \Delta_k$, the lowest risk of a natural estimator knowing p is

$$\tilde{r}_n^\ell(p, Q_{\text{nat}}) := \min_{\hat{p}' \in Q_{\text{nat}}} r_n^\ell(p, \hat{p}'),$$

and the *excess risk* of an arbitrary estimator $\hat{p}$ is

$$\tilde{r}_n^\ell(p, \hat{p}) := r_n^\ell(p, \hat{p}) - \tilde{r}_n^\ell(p, Q_{\text{nat}}).$$

Therefore, the worst-case excess risk, or *competitive risk*, of the estimator $\hat{p}$ over all distribution in Δ_k is

$$\tilde{r}_n^\ell(\hat{p}) := \max_{p \in \Delta_k} \tilde{r}_n^\ell(p, \hat{p}).$$

This formulation was introduced in [OS15] who showed that a simple variant of the Good-Turing estimator $\hat{p}_{\text{GT}}$ achieves a vanishing competitive KL-risk of $\tilde{r}_n(\hat{p}_{\text{GT}}) \leq (3+o(1))/n^{1/3}$, regardless of the alphabet size, and a more involved estimator $\hat{p}_{\text{MI}}$ achieves $\widetilde{\Theta}(\min\{k/n, 1/\sqrt{n}\})$, For ℓ_1-distance, [VV16] designed a linear-programming-based estimator $\hat{p}_{\text{LP}}$ and proved $\tilde{r}_n^{\ell_1}(\hat{p}_{\text{LP}}) = O(1/\text{polylog}(n))$.

Drawbacks The upper bounds provided by the competitive approach apply to all distributions, and similar to the min-max approach, track the excess error of the worst distribution. As we now show, they are too lax for many practical distributions. Consider the following generalization of the ubiquitous power-law distributions. For $c > 0$, $\alpha > 1$, and large alphabet-size k, define the enveloped distribution collection $\mathcal{P}_k^{\alpha,c} := \{p \in \Delta_k : p_x \leq c \cdot x^{-\alpha}\}$. It can be shown that for $n \in [k^{0.1}, k^2]$ there is a constant $C_{\alpha,c}$ depending on α and c, such that the min-max KL-risk of $\mathcal{P}_k^{\alpha,c}$ satisfies

$$r_n(\mathcal{P}_k^{\alpha,c}) = C_{\alpha,c} \cdot n^{-\frac{\alpha-1}{\alpha}+o(1)}.$$

By simple algebra, for $\alpha > 2$ and large n, this term is smaller than $\tilde{\Theta}(\min\{k/n, 1/\sqrt{n}\})$, the lowest competitive risk of any estimator [OS15]. Hence the guarantees the competitive framework provides do not suffice to address relatively "simple" common distributions.

3.2 New Results

The foregoing section reviewed the merits and drawbacks of classical and modern approaches to distribution-estimation. It noted that the min-max approach is "pessimistic" and often performs sub-optimally in both theory and practice. Of the modern frameworks, the structural approach works well if the structural assumptions are both correct and accurate, but fails otherwise, hence this approach is "local" but not "global". The competitive approach constructs universally near-optimal estimators, but provides the same guarantees regardless of the distribution's structure, potentially resulting in sub-optimal estimators for practical distributions, hence this approach is

"global" but not "local".

This raises the question of whether a single estimator can be both "global" and "local". Namely, without any assumptions on the distribution, provide universal excess-loss guarantees for general distributions, and stronger excess-loss guarantees for simple distributions. For example, an estimator $\hat{p}$ such that for any distribution p, $\tilde{r}_n(p,\hat{p}) \le n^{-1/2}$, and yet if the distribution p happens to be in the enveloped power-law class $\mathcal{P}_k^{3,c}$, then $\tilde{r}_n(p,\hat{p}) \le n^{-3/4}$.

We answer this question in the affirmative, and present the first competitive *and* structural distribution estimator.

3.2.1 Definitions

Instant competitive loss For consistency, let us instantiate the loss ℓ as the KL-divergence, i.e., for $p,q \in \Delta_k$,

$$\ell(p,q) := D(p \parallel q).$$

Let $p \in \Delta_k$ be an unknown discrete distribution, and let x^n be a realization of $X^n \sim p$. The best natural estimator, knowing both p and x^n, incurs the minimal possible loss

$$\tilde{\ell}_{x^n}(p, Q_{\text{nat}}) := \min_{\hat{p}' \in Q_{\text{nat}}} \ell(p, \hat{p}'(x^n)),$$

and for this particular pair (p,x^n), the excess loss of an arbitrary estimator $\hat{p}$ is

$$\tilde{\ell}_{x^n}(p,\hat{p}) := \ell(p,\hat{p}(x^n)) - \tilde{\ell}_{x^n}(p, Q_{\text{nat}}).$$

Hence for sequence x^n, the worst-case excess loss of $\hat{p}$ over Δ_k, or simply the *instance competitive loss* of $\hat{p}$, is

$$\tilde{\ell}_{x^n}(\hat{p}) := \max_{p \in \Delta_k} \tilde{\ell}_{x^n}(p,\hat{p}).$$

Permutation class For any distribution $p \in \Delta_k$, we denote by $\langle p \rangle$ the collection of

distributions in Δ_k that are equal to p up to some permutation over $[k]$. Knowing $\langle p \rangle$ is equivalent to knowing the multiset of p but not p itself.

General notation For $X^n \sim p \in \Delta_k$, the *multiplicity* of a symbol $x \in [k]$ is $N_x := \sum_{i=1}^{n} \mathbb{1}_{X_i=x}$, the number of times x appears in X^n. The *prevalence* of an integer μ is $\Phi_\mu := \sum_{x\in[k]} \mathbb{1}_{N_x=\mu}$, the number of symbols that appear μ times. Let $D := \sum_{\mu>0} \Phi_\mu$ be the number of distinct symbols in X^n, and let $D_\Phi := \sum_{\mu>0} \mathbb{1}_{\Phi_\mu>0}$ be the number of distinct positive multiplicities. Clearly, $D \geq D_\Phi$, and typically, $D \gg D_\Phi$. For example, if all symbols in the sequence X^n are distinct, then $D = n$, while D_Φ is just 1.

3.2.2 Main Results

We construct an explicit, near-linear-time computable distribution estimator $\hat{p}^*$ such that

Theorem 1. *For any distribution p, let $X^n \sim p$, then with probability at least $1 - n^{-8}$,*

$$\tilde{\ell}_{X^n}(\hat{p}^*) \leq \widetilde{O}\left(\frac{D_\Phi}{n}\right).$$

Note that the right-hand side is determined by just X^n, its computation requires no additional information about p.

The exact form of $\hat{p}^*$ can be found in Section 3.5, and the proof of Theorem 1 appears in the supplemental material.

Our main theorem implies the following new results and improvements on existing ones.

Global competitiveness In Section 3.3 we show that our estimator provides stronger estimation guarantees than many existing estimators: adaptive estimators (Corollary 2) such as the robust absolute discounting estimator [OD12, BHBO17]; competitive estimators (Corollary 4) such as the modified Good-Turing estimator [OS15]; and min-max estimators (Corollary 5).

Example: Section 3.3 shows that $D_\Phi \leq \min\{\sqrt{2n}, k\}$. Corollary 4 then concludes that the excess loss $\tilde{\ell}_{X^n}(\hat{p}^*)$ is always at most $\widetilde{O}(\min\{\sqrt{n}, k\}/n)$, providing a guarantee not only stronger

than the $n^{-1/3}$ rate of the modified Good-Turing estimator, but also as strong as the more involved estimator in [AJOS13b, OS15].

Local competitiveness In Section 3.4, we use the theorem to establish eight new results on learning important structured distributions. We show that our estimator has strong excess-loss bounds for three important structured distribution families: T-value (Corollary 7 and 8), log-concave (Corollary 9 and 10), and log-convex (Corollary 11, 12, and 13). Many common distributions are covered by these three classes. In particular, our results for power-law distributions (Corollary 12) are uniformly stronger than those in [FOOP17] for all parameter regimes.

Example: Corollary 8 shows that for all uniform distributions, $\mathbb{E}[D_\Phi]$ is bounded above by $\tilde{O}(n^{1/3})$, hence the algorithm's excess risk is at most $\tilde{O}(n^{-2/3})$.

Robustness to model misspecification The structural approach often uses different estimators for different distribution classes. By contrast, our single estimator provides robust and adaptive guarantees for a variety of structural classes without any modification.

Example: Over uniform distributions, $\hat{p}^*$ achieves an excess risk of $\tilde{O}(n^{-2/3})$ (Corollary 8), while for power law distributions with power parameter 1.5 , the same estimator achieves an excess risk of $\tilde{O}(n^{-3/5})$ (Corollary 11).

Robustness to domain permutations The structural approach often assumes that we know how to order the symbols so that the underlying distribution would exhibit certain structure (such as power-law). As discussed in Section 3.4, this assumption may be impractical. By contrast, since the distribution of D_Φ is the same for all $p' \in \langle p \rangle$, the excess loss/risk guarantees of our algorithm are invariant under any permutation of the domain symbols.

Example: If under some unknown ordering of the domain symbols, the underlying distribution is a power-law with power parameter 1.5 , then Corollary 11 implies that our estimator achieves an excess risk of $\tilde{O}(n^{-3/5})$.

Outline Besides Section 3.3 and 3.4 mentioned above, we present the exact form of our

estimator in Section 3.5.

3.3 Global Competitiveness

In this section, we present several implications of Theorem 1 for the universal estimation guarantees of $\hat{p}^*$. In particular, we show that $\hat{p}^*$ is near-optimal under various classical and modern distribution learning frameworks, including min-max and competitive mentioned above.

Corollary 1. *For any distribution p,*

$$\tilde{r}_n(p, \hat{p}^*) \leq \widetilde{O}\left(\frac{\mathbb{E}[D_\Phi]}{n}\right).$$

As in the proof of Theorem 1, $\tilde{\ell}_{X^n}(p, \hat{p}^*) \leq O(\log n)$ always. The corollary then follows from Theorem 1 itself.

Analogous to the previous definition of competitive distribution estimation, we can consider competing with an estimator that knows the probability multi-set. Specifically, for any distribution $p \in \Delta_k$, the lowest worst-case risk of a natural estimator knowing the multi-set of p is

$$\dot{r}_n^\ell(\langle p \rangle) := \min_{\hat{p}'} \max_{p' \in \langle p \rangle} r_n^\ell(p', \hat{p}'),$$

and an arbitrary estimator $\hat{p}$ has the *multi-set excess risk* of

$$\dot{r}_n^\ell(p, \hat{p}) := r_n^\ell(p, \hat{p}) - \dot{r}_n^\ell(\langle p \rangle).$$

For KL-divergence, the following lemma relates $\dot{r}_n^\ell$ to $\tilde{r}_n^\ell$.

Lemma 1. *[OS15] For any distribution $p \in \Delta_k$ and estimator $\hat{p}$,*

$$\max_{p' \in \langle p \rangle} \dot{r}_n(p', \hat{p}) \leq \tilde{r}_n(p, \hat{p}).$$

Together with Corollary 1, the lemma yields,

Corollary 2. *For any distribution p,*

$$\max_{p' \in \langle p \rangle} \dot{r}_n(p, \hat{p}^*) \leq \widetilde{O}\left(\frac{\mathbb{E}[D_\Phi]}{n}\right).$$

Adaptive optimality The min-max results [KT81] imply that for any estimator, learning an arbitrary k-symbol distribution up to a certain KL-risk requires $\Omega(k)$ samples in the worst case. Since modern data science often considers applications over large alphabets, this is normally viewed as a negative result. However, as experience suggests, many practical distributions have small "effective alphabet sizes". For example, if we draw 10 samples from a geometric distribution with success probability 0.9, although the support size is infinite, with high probability, we shall observe at most 3 distinct symbols.

To formalize this intuition, for a given n, let the *effective alphabet size* of a distribution p be the expected number $\mathbb{E}[D]$ of distinct symbols that appear in $X^n \sim p$. As in [FOOP17], given n, k, and d, let $\mathcal{P}_d$ be the collection of distributions in Δ_k satisfying $\mathbb{E}[D] \leq d$. By Corollary 1, the performance of $\hat{p}^*$ over $\mathcal{P}_d$ is adaptive to d:

Corollary 3. *For all $d \geq 2$ and every distribution $p \in \mathcal{P}_d$,*

$$r_n(p, \hat{p}^*) \leq \frac{d}{n} \log k + \widetilde{O}\left(\frac{d}{n}\right).$$

The following lemma shows the optimality of Corollary 3.

Lemma 2. *[FOOP17] Let α be any constant greater than 1. There exist constants $c_0 > 0$ and n_0 such that for $d = n^{\frac{1}{\alpha}}$, any estimator $\hat{p}$, all $n > n_0$, and all $k > \max\{3n, 1.2^{\frac{1}{\alpha-1}} n^{\frac{1}{\alpha}}\}$,*

$$\max_{p \in \mathcal{P}_d} r_n(p, \hat{p}) \geq c_0 \frac{d}{n} \log k - \widetilde{O}\left(\frac{d}{n}\right).$$

Here we present two immediate implications. First, to learn a k-symbol distribution up to a certain KL-risk, the number of samples we need is at most $\tilde{O}(\mathbb{E}[D]\log k)$, which is often much smaller than $\Omega(k)$. Second, in the extreme case when $k/n \to \infty$, the upper bound on $r_n(p,\hat{p}^*)$ is at most $(1+o(1))\log k$. Hence, our estimator achieves the min-max KL-risk over Δ_k to the right constant.

Competitive optimality Now we show that $\hat{p}^*$ is near-optimal under the competitive formulation described in Section 3.1.2. We begin by finding a simple upper bound for D_Φ, the number of distinct positive multiplicities. Since different multiplicities correspond to distinct symbols, D_Φ is at most the alphabet size k. On the other hand, since only distinct positive multiplicities count, $\sum_{\mu=1}^{D_\Phi}\mu \le n$. Hence, $D_\Phi \le \min\{k,\sqrt{2n}\}$, which together with Corollary 1 yields

Corollary 4. *For any distribution p,*

$$\tilde{r}_n(p,\hat{p}^*) \le \widetilde{O}\left(\frac{\min\{k,\sqrt{n}\}}{n}\right).$$

The following lemma shows the optimality of Corollary 4.

Lemma 3. *[OS15] For any estimator $\hat{p}$,*

$$\max_{p\in\Delta_k}\tilde{r}_n(p,\hat{p}) \ge \widetilde{\Omega}\left(\frac{\min\{k,\sqrt{n}\}}{n}\right).$$

Min-max optimality The previous results show that $\hat{p}^*$ often achieves the min-max KL-risk $r_n(\Delta_k)$ to the right constant. Specifically,

Corollary 5. *Let α_0 be any constant greater than $1/2$. For any $\alpha > \alpha_0$ and $k > n^\alpha$,*

$$r_n(\Delta_k,\hat{p}^*) = (1+o_n(1))r_n(\Delta_k).$$

3.4 Local Competitiveness

We use Corollary 1 and 3 to establish eight new results on learning important structured distributions. We show that our estimator has strong excess-loss bounds for three important structured distribution families: T-value (Corollary 7 and 8), log-concave (Corollary 9 and 10), and log-convex (Corollary 11, 12, and 13). Many common distributions are covered by these three classes.

3.4.1 A Simple Bound on $\mathbb{E}[D_\Phi]$

By Corollary 1, the excess KL-risk $\tilde{r}_n(p,\hat{p}^*)$ of $\hat{p}^*$ in estimating p is upper bounded by $\widetilde{O}(\mathbb{E}[D_\Phi]/n)$. Perhaps the most natural question to ask is: given n and p, how large is $\mathbb{E}[D_\Phi]$? To get a relatively simple closed-form expression for $\mathbb{E}[D_\Phi]$, we adopt the conventional "Poisson Sampling" technique where the sample size is an independent Poisson variable with mean n. By doing so, the multiplicities $N_x \sim \text{Poi}(np_x)$ independently of each other. Under Poisson sampling, the linearity of expectation implies

$$\mathbb{E}[D_\Phi] = n - \sum_{\mu>0} \prod_{x\in[k]} \left(1 - e^{-np_x}\frac{(np_x)^\mu}{\mu!}\right).$$

Expanding the right-hand side would give us an expression consisting of $n\cdot(2^k-1)$ terms, which is hard to analyze. Hence, instead of evaluating $\mathbb{E}[D_\Phi]$ directly, we would like to work on its simple upper bounds. Given sampling parameter n, we partition the unit-length interval $(0,1]$ into a sequence of sub-intervals,

$$I_j := \left((j-1)^2\frac{\log n}{n}, j^2\frac{\log n}{n}\right], \ 1 \le j \le \sqrt{\frac{n}{\log n}}.$$

For any distribution p, denote by p_{I_j} the number of probabilities p_x in I_j. Then,

Lemma 4. *For any distribution p,*

$$\mathbb{E}[D_\Phi] \leq O(\sum_{j\geq 1} \min\left\{p_{I_j}, j\right\}) \cdot \log n.$$

In addition, since p is a distribution, for all j, $p_{I_j} \cdot \frac{j^2 \log n}{n} \leq 1$, which in turn implies

$$\min\left\{p_{I_j}, j\right\} \leq \min\left\{\frac{n}{j^2 \log n}, j\right\} < n^{\frac{1}{3}}.$$

More generally, let P_{I_j} denote the *sum* of probabilities p_x in I_j. Then,

$$\min\left\{p_{I_j}, j\right\} \leq \min\left\{\frac{nP_{I_j}}{j^2 \log n}, j\right\} < (nP_{I_j})^{\frac{1}{3}}.$$

Combined, Corollary 1 and Lemma 4 yield

Corollary 6. *For any distribution p,*

$$\tilde{r}_n(p, \hat{p}^*) \leq \widetilde{O}\left(\frac{1}{n}\right) \sum_{j\geq 1} \min\left\{p_{I_j}, j\right\}.$$

To illustrate the Corollary's significance, we present its implications for various distribution learning problems.

3.4.2 T-Value Distributions

A uniform distribution can be described as a distribution whose positive probabilities take only a single value. As a generalization of this formulation, we call a distribution p a *T-value distribution* if its positive probabilities p_x can take T different values. Note that T-value distributions over $[k]$ can be viewed as mixtures of T uniform distributions over different subsets of $[k]$, and that these distributions generalize T-piecewise histogram distributions. Intuitively, for smaller values of T, we would expect the task of learning an unknown T-value distribution to be

easier. The following corollary confirms this intuition.

Corollary 7. *For any T-value distribution p and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}\left(\frac{T^{\frac{2}{3}} \wedge n^{\frac{1}{6}}}{n^{\frac{2}{3}}}\right).$$

Note that $p \in \langle p \rangle$. To prove the corollary, observe that by our previous result, for all j,

$$\min\left\{p_{I_j}, j\right\} < (nP_{I_j})^{\frac{1}{3}}.$$

Note that for a T-value distribution, $p_{I_j} \neq 0$ for at most T different j values, say $j_1, \ldots, j_T$. By the above inequality and Corollary 6,

$$\tilde{r}_n(p, \hat{p}^*) \leq \widetilde{O}\left(\frac{1}{n}\right) \sum_{i=1}^{T} (nP_{I_{j_i}})^{\frac{1}{3}} \leq \widetilde{O}\left(\frac{T(n/T)^{\frac{1}{3}}}{n}\right).$$

combined with Corollary 4, this completes the proof.

Uniform Distributions Now we consider the collection $\mathcal{U}_k$ of 1-value distributions, i.e., uniform distributions over non-empty subsets of $[k]$. Our objective is to derive a result stronger than Corollary 7. Let S_p denote the support size of a distribution $p \in \mathcal{U}_k$. For all $x \in [k]$, p_x is either 0 or S_p^{-1}. Since $\{I_j, j \geq 1\}$ is a partition of $(0, 1]$, there exists a unique j' such that $S_p^{-1} \in I_{j'}$, i.e.,

$$S_p^{-1} \in I_{j'} = \frac{\log n}{n}\left((j'-1)^2, j'^2\right],$$

further implying $1 + \sqrt{n/(S_p \log n)} \geq j'$. Together with $D_\Phi \leq D \leq S_p$ and Corollary 6, this shows

Corollary 8. *Let p be an arbitrary distribution in $\mathcal{U}_k$, then*

$$\tilde{r}_n(p, \hat{p}^*) \leq \tilde{O}\left(\min\left\{\frac{1}{\sqrt{nS_p}}, \frac{S_p}{n}\right\}\right).$$

Note that the right-hand side is no more than $\tilde{O}(n^{-2/3})$. Furthermore, in both the small alphabet regime where $S_p = O(1)$ and the large alphabet regime where $S_p = \Omega(n)$, we have $\tilde{r}_n(p, \hat{p}^*) \leq \tilde{O}(n^{-1})$, which is fairly tight.

3.4.3 Log-Concave Distributions

The class of discrete log-concave distributions covers a variety of well-known distribution classes including binomial, Poisson, negative binomial, geometric, hypergeometric, hyper-Poisson, Skellam, and Pólya-Eggenberger [QBW90]. We say a discrete distribution $p \in \Delta_k$ is *log-concave* if for all $x \in [k]$, $p_x^2 \geq p_{x-1} \cdot p_{x+1}$, and denote the collection of all such distributions by $\mathcal{L}_k$. Further, for all $\sigma > 0$, let $\mathcal{L}_k^{n,\sigma}$ denote the collection of $p \in \mathcal{L}_k$ whose standard deviation lies in $(\sigma \cdot \log^{-1} n, \sigma]$. Intuitively, one would expect the learning task over $\mathcal{L}_k^{n,\sigma}$ to be easier for smaller values of σ. The following corollary demonstrates the correctness of this intuition and shows the competitive performance of our estimator. Due to space considerations, we postpone its proofs to the supplemental material.

Corollary 9. *For any distribution $p \in \mathcal{L}_k^{n,\sigma}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}\left((\sigma n)^{-\frac{1}{3}} \wedge \frac{\sigma}{n}\right).$$

For any $\sigma \gg 1$, the right-hand side is uniformly smaller than the bound $\tilde{O}(\min\{k, \sqrt{n}\} \cdot n^{-1})$ in Corollary 4.

For mixtures of distributions in $\mathcal{L}_k^{n,\sigma}$, an analogous argument gives the following result.

Corollary 10. *Let p be a t-mixture of distributions in $\mathcal{L}_k^{n,\sigma}$ and p' be any distribution in $\langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}\left((\sigma n)^{-\frac{1}{3}} \wedge \frac{t\sigma \wedge \sqrt{n}}{n}\right).$$

3.4.4 Log-Convex Distributions

While the T-value and log-concave families cover many common distributions, there are certainly more distribution classes to be explored. For example, a truncated power-law distribution is always log-convex. In this section, we consider two generic classes of log-convex distributions: power-law and Hurwitz–Lerch Zeta distribution families.

Enveloped power-law distributions Consider the collection $\mathcal{P}_k^{\alpha,c} := \{p \in \Delta_k : p_x \leq c \cdot x^{-\alpha}\}$ of enveloped (truncated) power-law distributions. Note that this definition generalizes power-law families, and that distributions in $\mathcal{P}_k^{\alpha,c}$ are not necessarily log-convex. We have the following result, whose proof appears in the supplemental material.

Corollary 11. *For any distribution $p \in \mathcal{P}_k^{\alpha,c}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}_{c,\alpha}\left(n^{-\max\{\frac{\alpha}{\alpha+1}, \frac{1}{2}\}}\right).$$

The distribution collection $\mathcal{P}_k^{\alpha,c}$ has the interesting property that it is closed under mixtures. Hence, Corollary 11 also covers mixtures of enveloped power-law distributions.

Implications of Corollary 11 Let $p^\alpha \in \Delta_k$ be the truncated power-law distribution with power α that is truncated at k, i.e., $p_x^\alpha \propto x^{-\alpha}$, $\forall x \in [k]$. Clearly, we have $p^\alpha \in \mathcal{P}_k^{\alpha,c}$ for all $c \geq 1$. The recent work of [FOOP17] shows that for $k > \{n, n^{\frac{1}{\alpha-1}}\}$ and any distribution $p' \in \langle p^\alpha \rangle$, the estimator $\hat{p}''$ proposed in [OD12] satisfies

$$\dot{r}_n(p', \hat{p}'') \leq O_{c,\alpha}\left(n^{-\frac{2\alpha-1}{2\alpha+1}}\right).$$

A simple combination of Lemma 1 and Corollary 11 yields

Corollary 12. *For any distribution $p' \in \langle p^\alpha \rangle$,*

$$\dot{r}_n(p', \hat{p}^*) \leq \tilde{r}_n(p, \hat{p}^*) \leq \tilde{O}_{c,\alpha}\left(n^{-\max\{\frac{\alpha}{\alpha+1}, \frac{1}{2}\}}\right).$$

Our approach has the following three advantages over the previous result in [FOOP17]. First, for all $\alpha > 0$, we have $-\alpha/(\alpha+1) < -(2\alpha-1)/(2\alpha+1)$, hence our guarantee is uniformly better than the previous one. Second, the previous result requires $k > \{n, n^{\frac{1}{\alpha-1}}\}$ to hold, which can be non-realistic for α close to 1. In comparison, our result does not require such conditions at all. Third, for small $\alpha < 1/2$, the previous result only implies a multi-set excess risk of $O(n^{\Theta(1)})$, while Corollary 12 always yields $\tilde{O}\left(n^{-1/2}\right)$ regardless of α.

Enveloped Hurwitz–Lerch Zeta distributions For any distribution $p \in \Delta_k$, p is a (truncated) Hurwitz–Lerch Zeta (HLZ) distribution [GGOS08] if

$$p_x = \frac{1}{T(\theta, s, a, k)} \cdot \frac{\theta^x}{(a+x)^{s+1}},$$

for some parameter $s \geq 0$, $a \in [0,1]$ and $\theta \in (0,1]$, where the normalization factor $T(\theta, s, a, k) := \sum_{x\in[k]} \theta^x/(a+x)^{s+1}$. Analogously, consider $\mathcal{H}_k^{\theta,s,a,c} := \{p \in \Delta_k : p_x \leq c \cdot \theta^x/(a+x)^{s+1}\}$ of *enveloped* HLZ distributions. HLZ distributions include the well-known Riemann Zeta, Zipf-Mandelbrot, Lotka, Good, logarithmic-series, and Estoup distributions. These distributions have various applications in many fields. For example, the Good distribution [ZA95] can be used to model species' frequencies and to estimate population parameters.

Note that $\mathcal{H}_k^{\theta,s,a,c} \subseteq \mathcal{P}_k^{s+1,c}$ for $\alpha \geq 1$. Hence, by Corollary 11, for any distribution $p \in \mathcal{H}_k^{\theta,s,a,c}$ and $p' \in \langle p \rangle$,

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}_{c,s}\left(n^{-\frac{s+1}{s+2}}\right).$$

Let x_1 be the threshold parameter such that $c \cdot \theta^{x_1} = n^{-1}$. Direct computation gives $x_1 = \log(cn)/\log\frac{1}{\theta}$. The symbols $x \in [k]$ that are no larger than x_1 contribute at most x_1 to $\mathbb{E}[D_\Phi]$. Furthermore, the proof of Lemma 4 essentially shows that symbols with probability no larger than n^{-1} contributes at most $O(\log n)$ to $\mathbb{E}[D_\Phi]$. We conclude that $\mathbb{E}[D_\Phi] \leq O(\log(cn)/\log\frac{1}{\theta} + \log n)$.

Corollary 3 combines the above results and yields

Corollary 13. *For any $p \in \mathcal{H}_k^{\theta,s,a,c}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}_{c,s}\left(\frac{n^{\frac{1}{s+2}}}{n} \wedge \frac{1-\log^{-1}\theta}{n}\right).$$

Note that the right-hand side is the minimum of two quantities. For $\theta \in (1-n^{-\frac{1}{s+2}}, 1]$, we can reduce the upper bound to $\tilde{O}_{c,s}(n^{-\frac{s+1}{s+2}})$. On the other hand, for $\theta \in (0, 1-n^{-\frac{1}{s+2}}]$, the upper bound becomes $\tilde{O}_{c,s}((1-\log^{-1}\theta)\cdot n^{-1})$.

3.4.5 Robustness to Domain Permutations

Our results on learning structured distribution families differ significantly from nearly all the existing ones. Prior work has mainly considered unknown distribution with a certain structure over a known and ordered domain. In our formulation, we assume that the underlying distribution has certain structure under some particular ordering of the domain elements, and this ordering is unknown to the estimator.

Below we illustrate this by a concrete example.

Let $\mathcal{F}$ be a finite discrete domain of size k. Consider learning an unknown log-concave distribution $P \in \Delta_{\mathcal{F}}$ from its sample sequence Y^n. Traditional formulations like [CDSS13] assume that we know an exact bijective mapping σ from $\mathcal{F}$ to $[k]$, such that reordering the probabilities of P according to σ yields a log-concave distribution $p \in \Delta_k$. Further applying σ to Y^n and denoting the resulting sequence by X^n transforms the problem into learning p from a sample sequence $X^n \sim p$. Here, the assumption that p is log-concave is equivalent to requiring $p_x^2 \geq p_{x-1} \cdot p_{x+1}$, for all $x \in [k] \setminus \{1, k\}$. We can see that such formulation may be non-practical. For example, in natural language processing, the observed samples are words and punctuation marks. Even we know these samples come from a log-concave distribution, we don't know how to order the alphabet, i.e., find the right mapping σ, so that the corresponding distribution $p \in \Delta_k$ would be log-concave.

3.5 The Estimator

Let p be an arbitrary distribution in Δ_k, and let X^n be a length-n sample sequence from p. For simplicity, abbreviate $\mathbb{1}_x^\mu := \mathbb{1}_{N_x=\mu}$. For any natural number μ, denote the total probability mass of the symbols that appear μ times by

$$M_\mu := \sum_{x\in[k]} p_x \mathbb{1}_x^\mu.$$

After observing X^n, an estimator $\hat{p}$ approximates M_μ by

$$\hat{M}_\mu := \sum_{x:N_x=\mu} \hat{p}_x(X^n).$$

Assume that $\hat{p}$ is a natural estimator. By [OS15], the excess loss of $\hat{p}$ over the best natural estimator that knows the underlying distribution p is

$$\tilde{\ell}_{X^n}(p,\hat{p}) = D(M \parallel \hat{M}) := \sum_{\mu\geq 0} M_\mu \log \frac{M_\mu}{\hat{M}_\mu}.$$

The above characterization of $\tilde{\ell}_{X^n}(p,\hat{p})$ converts the problem of finding good natural estimators for the underlying distribution to that of finding good estimators for

$$M := (M_0,\ldots,M_n).$$

Intuition We first motivate the estimator, whose form is similar to that in [AJOS13b], but with some modifications. Since the estimator is natural, it needs to approximate only $M := (M_0,\ldots,M_n)$. The construction is guided by analyzing the estimator bias and concentration properties for various multiplicities μ. To estimate M_0, we use the provably near-optimal [RTS17] Good-Turing estimator. For the remaining multiplicities, analysis shows that for moderate, yet

frequent multiplicities, namely $\mu = O(\log n)$ and $\Phi_\mu = \Omega(\log^2 n)$, the Good-Turing estimator performs nearly optimally. For infrequent multiplicities, the empirical estimator performs better. For the remaining multiplicities, both estimates are sub-optimal. Applying polynomial approximation techniques, we construct a more involved estimator that approximates the behavior of a genie that knows that expected M_μ values. The estimator is slightly simpler than that in [AJOS13b], yet achieves better performance.

Details Since our estimator $\hat{p}^*$ is natural, we simply specify $\hat{M}^*_\mu := \sum_{x:N_x=\mu} \hat{p}^*_x(X^n)$. To simplify the analysis, we adopt the standard "Poisson sampling" technique, and make the sample size a Poisson variable N with mean value n.

For $N < n\log n$, let c_1, c_2, and c_3 be properly chosen absolute constants. For any two natural numbers $\mu \geq \mu'$, denote $a^{\mu'}_\mu := \mu'!/\mu!$ and $E^{\mu'}_{x,\mu} := \mathbb{1}^{\mu'}_x a^{\mu'}_\mu (N_x)^{\underline{\mu-\mu'}}$, where $A^{\underline{B}}$ is the falling factorial of A of order B. Let

$$E_{x,\mu} = \frac{1}{c_1\sqrt{\mu/\log n}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E^{\mu'}_{x,\mu}.$$

We can show that $E_\mu := \sum_{x\in[k]} E_{x,\mu}$ is an unbiased estimator of $\mathbb{E}[\Phi_\mu]$. Empirical-frequency estimates M_μ by

$$\hat{\phi}_\mu := \Phi_\mu \frac{\mu}{n},$$

while Good-Turing estimates it by

$$\hat{G}_\mu := \Phi_{\mu+1} \frac{\mu+1}{n}.$$

To avoid zero probability estimates, we modify the Good-Turing estimator to $\hat{G}'_\mu := \max\{1/n, \hat{G}_\mu\}$ and let

$$\hat{O}_\mu := \Phi_\mu \frac{\mu+1}{n} \frac{E_{\mu+1}}{E_\mu},$$

and similarly set

$$\hat{O}'_\mu := \min\{\max\{1/n, \hat{O}_\mu\}, \log^2 n\}.$$

For $\mu < n \log n$, our estimator is

$$\hat{M}^*_\mu = \begin{cases} \hat{G}'_\mu & \text{if } \mu = 0, \\ \hat{\phi}_\mu & \text{if } \mu \geq 1 \text{ and } \Phi_\mu \leq c_2(\log^2 n), \\ \hat{O}'_\mu & \text{if } \mu > c_3 \log n \text{ and } \Phi_\mu > c_2(\log^2 n), \\ \hat{G}'_\mu & \text{if } c_3 \log n \geq \mu \geq 1 \text{ and } \Phi_\mu > c_2(\log^2 n). \end{cases}$$

As Poisson variables are concentrated around their mean, for $N \geq n \log n$, which rarely happens, and $\mu \in [0, N]$, we simply set $\hat{M}^*_\mu = 1/(N+1)$. If these probability estimates do not sum to 1, we normalize them by their sum.

Finally for each $x \in [k]$, our distribution estimator is

$$\hat{p}^*_x(X^n) = \frac{\hat{M}^*_{N_x}}{\Phi_{N_x}}.$$

3.6 Numerical Experiments

The estimator is easy to implement. In Section 1 of the supplemental material, we present experimental results on a variety of distributions, and show that the proposed estimator indeed outperforms the improved Good-Turing estimator in [OS15].

3.7 Future Directions

The results obtained in paper strengthen and extend the competitive approach to distribution estimation taken in [OS15]. It would be of interest to obtain similar results for distribution

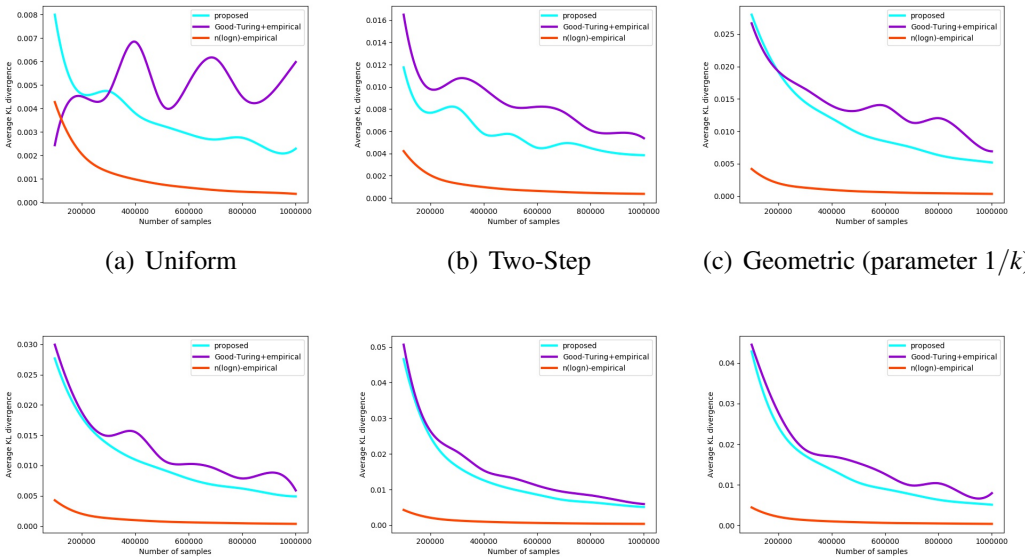
estimation under ℓ_1 distance. [KOPS15] showed that the simple empirical estimator achieves the min-max ℓ_1-risk $r_n^{\ell_1}(\Delta_k) = (1+o(1))\sqrt{2(k-1)/(\pi n)}$. Yet the excess risk of the estimator in the nice work of [VV16] is $O(1/\text{polylog}(n))$. Hence, for $k \leq \tilde{O}(n)$, this guarantee does not improve that of the empirical estimator, raising the possibility of strengthening the competitive results.

A similar approach can be applied to the related *property-estimation* task. A property, e.g., Shannon entropy, is simply a mapping $f : \Delta_k \to \mathbb{R}$. Most existing property-estimation results are worst-case (min-max) in nature. Yet practical and natural distributions are rarely the worst possible, and often possess a simple structure. To address this discrepancy, recent works [HOSW18, HO20a] took a competitive approach, constructing estimators whose performance is adaptive to the simplicity of the underlying distribution. Specifically, the widely-used empirical estimator estimates property values by evaluating the property at the empirical distribution. For every property in a broad class and *every* distribution in Δ_k, the expected error of the estimator in [HO20a] with sample size $n/\log n$ is at most that of the empirical estimator with sample size n, plus a distribution-free vanishing function of n.

These results cover several well-known properties such as entropy and support size, for which the $\log n$ factor is optimal up to constants, and also apply to any property in the form of $\sum_x f_x(p_x)$, such as the ℓ_1 distance to a given distribution, where f_x is 1-Lipschitz for all $x \in [k]$. It would be of interest to construct a doubly-competitive estimator for property estimation as well.

3.8 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in *International Conference on Machine Learning*. Yi Hao and Alon Orlitsky, 2019. The dissertation author was the primary investigator and author of this paper.



(a) Uniform (b) Two-Step (c) Geometric (parameter $1/k$)

(d) Zipf with parameter 0.5 (e) Uniform prior (Dirichlet 1) (f) Dirichlet 2 prior

Figure 3.1: Experimental results for support $k = 10,000$, number of samples n ranging from $10k$ to $100k$, averaged over 30 independent trials.

3.A Experiments

Experimental plots and relevant details are shown below.

Estimators We consider three estimators: the proposed estimator with sample size n, the improved Good-Turing estimator [OS15] with the same sample size, and the empirical estimator with a larger $n \log n$ sample size. As shown in [OS15], the improved Good-Turing estimator considerably outperforms other estimators such as the Laplace estimator (add-1 estimator), the Krichevsky-Trofimov estimator [KT81], and the Braess-Sauer estimator [BS04]. Hence we do not include the latter estimators here.

Hyper-Parameters Our algorithm employs three hyper-parameters: c_1 is inversely related to the variance of the probability estimates and is best chosen above 1, c_2 controls the boundary between frequent and infrequent multiplicities and is best chosen below 1, and c_3 is proportional to the threshold separating small and large probabilities and is best chosen be around 1. In the experiments, we simply set $c_1 = 2$, $c_2 = 0.5$, and $c_3 = 1$.

Distributions We choose alphabet size $k = 10,000$ and consider six different distributions over $[k]$: a uniform distribution of support size k; a two-step distribution with half the symbols having probability $1/(2k)$, and the other half having probability $3/(2k)$; a geometric distribution with parameter $g = 1/k$, i.e., $p_i = (1-g)^{i-1}g$, truncated at $i = k$ and renormalized; a Zipf distribution with parameter 0.5, i.e., $p_i \propto i^{-0.5}$, truncated at $i = k$ and renormalized; a distribution generated by the uniform prior on Δ_k; and a distribution generated by a Dirichlet-2 prior.

Experimental settings For each distribution we repeated the experiments 30 times and show the average KL-divergence between the underlying distribution and the distribution estimates. The relative performance of the three estimators is consistent over a wide range of sample sizes. To better differentiate the performance of the three estimators, we limit the dynamic range of the error by showing the results for sample sizes n ranging from $10 \cdot k$ to $100 \cdot k$.

Code The code is available at `https://github.com/ucsdyi/Competitive`.

Conclusions As can be observed in all six plots, the proposed estimator outperforms the improved Good-Turing estimator. Because of the estimator construction, outlined in Section 5 of the main paper, the improvement is most pronounced when $n \geq k$.

3.B Proof of Theorem 1

In this section we prove Theorem 1 in the main paper.

Proof sketch From the discussion in Section 5 of the main paper, we need to estimate only M_μ. Relations such as $\mathbb{E}[M_{\mu-1}] = (\mu/n)\mathbb{E}\left[\Phi_\mu\right]$ suggest constructing estimators for $\mathbb{E}\left[\Phi_\mu\right]$. By the identity $\mathbb{E}\left[\Phi_\mu\right] = \sum_{x\in[k]} \mathbb{E}\left[\mathbb{1}_x^\mu\right]$, we can further reduce the problem to estimating $\mathbb{E}\left[\mathbb{1}_x^\mu\right]$. We then approximate $\mathbb{E}\left[\mathbb{1}_x^\mu\right]$ by scaled versions of $\mathbb{1}_x^{\mu'}$ where μ' is close to μ. This simple approach yields unbiased estimators $E_{x,\mu}^{\mu'}$ with sub-optimal variances. An important observation is that $\mathbb{1}_x^\mu \cdot \mathbb{1}_x^{\mu'} = 0$ for all $\mu \neq \mu'$, making it possible to construct a new estimator $E_{x,\mu}$ with

near-optimal variance by averaging a sequence of these unbiased estimators. Note that $E_{x,\mu}$ is still unbiased. Summing the estimators over $[k]$, we estimate $\mathbb{E}\left[\Phi_\mu\right]$ by $E_\mu := \sum_{x\in[k]} E_{x,\mu}$. As shown in [AJOS13b], a genie that knows both $\mathbb{E}[\Phi_{\mu+1}]$ and $\mathbb{E}[\Phi_\mu]$ could accurately estimate M_μ by $(\Phi_\mu(\mu+1)/n)(\mathbb{E}[\Phi_{\mu+1}]/\mathbb{E}[\Phi_\mu])$. Hence, to approximate the genie's performance, we leverage the estimator for $\mathbb{E}[\Phi_\mu]$ and use $\hat{O}_\mu := (\Phi_\mu(\mu+1)/n)(E_{\mu+1}/E_\mu)$. Note that this estimator is the ratio of two estimators and hence not easy to analyze. To simplify the analysis, we modify E_μ slightly so that it has a structure similar to that of $E_{\mu+1}$. Then we prove that for relatively large, and frequent multiplicities, namely $\mu = \Omega(\log n)$ and $\Phi_\mu = \Omega(\log^2 n)$, the proposed estimator almost achieves the performance of the genie. As illustrated in Section 5 of the main paper, for other multiplicities, analysis shows that Good-Turing and empirical estimators are already near-optimal. Combined, these estimates form our final estimator for the vector M, and establish the guarantees stated in Theorem 1.

The Expected Total Probability Mass

To simplify our analysis, we adopt the standard "Poisson sampling" technique [MU05]. Instead of having a sample sequence of fixed length n, we make the sample size a Poisson random variable N with mean value n. Let p be an arbitrary distribution over $[k]$, and X^N be a length-$\mathrm{Poi}(n)$ sample sequence from p. Let N_x denote the number of times symbol x appearing in X^N, and let Φ_μ denote the number of symbols appearing μ times. For simplicity, denote $\mathbb{1}_x^\mu := \mathbb{1}_{N_x=\mu}$. Then, the total probability mass of the symbols that appear μ times is

$$M_\mu = \sum_{x\in[k]} p_x \mathbb{1}_x^\mu.$$

By the argument in Section 5 of the main paper, it suffices to design an estimator for M_μ.

The expectation of M_μ is

$$\mathbb{E}[M_\mu] = \mathbb{E}\left[\sum_{x\in[k]} p_x \mathbb{1}_x^\mu\right] = \sum_{x\in[k]} p_x e^{-np_x}\frac{(np_x)^\mu}{\mu!} = \frac{\mu+1}{n}\sum_{x\in[k]}\mathbb{E}\left[\mathbb{1}_x^{\mu+1}\right] = \frac{\mu+1}{n}\mathbb{E}\left[\Phi_{\mu+1}\right].$$

Furthermore, as shown in [AJOS13b], a genie that knows both $\mathbb{E}[\Phi_{\mu+1}]$ and $\mathbb{E}[\Phi_\mu]$ could estimate M_μ really well using the estimator

$$O_\mu := \Phi_\mu \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}.$$

Both observations suggest that we should find a good estimator for $\mathbb{E}[\Phi_{\mu+1}]$.

Estimating an Indicator Variable

The above derivation shows that $\mathbb{E}\left[\Phi_{\mu+1}\right] = \sum_{x\in[k]}\mathbb{E}\left[\mathbb{1}_x^{\mu+1}\right]$. Symmetry further reduces the problem to estimating a single term $\mathbb{E}\left[\mathbb{1}_x^{\mu+1}\right]$. For notational convenience, we change $(\mu+1)$ to μ. For any two natural numbers μ and μ', let $a_\mu^{\mu'} := \mu'!/\mu!$. Direct computation yields

$$\mathbb{E}[\mathbb{1}_x^\mu] = \mathbb{E}[\mathbb{1}_x^{\mu'}]a_\mu^{\mu'}(np_x)^{\mu-\mu'}.$$

To further simplify our derivations, let us assume that another two independent length-$\mathrm{Poi}(n)$ sample sequences from p are given, say $X^{N'}$ and $X^{N''}$ where $N' \sim \mathrm{Poi}(n)$ and $N'' \sim \mathrm{Poi}(n)$. Denote by N_x' and N_x'' the number of times symbol x appearing, and Φ_μ' and Φ_μ'' the number of symbols appearing μ times, in $X^{N'}$ and $X^{N''}$, respectively. This is equivalent to the commonly-used "sample splitting" technique [WY16], namely, we split the given sample sequence into three independent subsequences of roughly the same length. It is not hard to see that even without these additional sample sequences, performing sample splitting shall change the right-hand side of Theorem 1 by at most a multiplicative factor of three, hence does not affect the statement of the theorem. By the last identity and properties of Poisson random variables, for $\mu \geq \mu'$, the

following estimator is an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^{\mu}]$,

$$E_{x,\mu}^{\mu'} := \mathbb{1}_x^{\mu'} a_{\mu}^{\mu'} (N'_x)^{\underline{\mu-\mu'}},$$

where $A^{\underline{B}}$ is the falling factorial of A of order B.

Let c_1 be a positive absolute constant. In the subsequent proofs, we will assume that c_1 is sufficiently small and lies in $(0,1)$ to avoid large constants in the expressions. For $c_1 > 1$, the proof of Theorem 1 still follows from the remaining arguments. Other related constants have also been chosen to simplify the proofs and expressions. For example, we set $c_3 = 100$ to eliminate some edge cases.

While the bias of $E_{x,\mu}^{\mu'}$ in estimating $\mathbb{E}[\mathbb{1}_x^{\mu}]$ is zero, the variance of $E_{x,\mu}^{\mu'}$ satisfies

$$\mathrm{Var}(E_{x,\mu}^{\mu'}) \leq \mathbb{E}[(E_{x,\mu}^{\mu'})^2] \leq \left(a_{\mu}^{\mu'}\right)^2 \mathbb{E}[\mathbb{1}_x^{\mu'}] \cdot \mathbb{E}\left[\left((N'_x)^{\underline{\mu-\mu'}}\right)^2\right].$$

The quantity on the right-hand side is the product of three terms. We bound the first term using the following lemma.

Lemma 5. *For sufficiently large n and any two natural numbers μ, μ' such that $n\log n > \mu > 100\log n$ and*

$$\mu - c_1\sqrt{\frac{\mu}{\log n}} \leq \mu' \leq \mu - 1,$$

we have

$$(a_{\mu}^{\mu'})^2 \leq \frac{4}{(\mu^{\mu-\mu'})^2}.$$

Proof. The quantity of interest satisfies

$$\begin{aligned}
a_{\mu}^{\mu'} &= \frac{\mu'!}{\mu!} \\
&= \frac{1}{\mu^{\mu-\mu'}} \cdot \frac{\mu^{\mu-\mu'}}{\prod_{j=0}^{\mu-\mu'-1}(\mu-j)} \\
&\leq \frac{1}{\mu^{\mu-\mu'}} \cdot \left(\frac{\mu}{\mu - c_1\sqrt{\frac{\mu}{\log n}}} \right)^{c_1\sqrt{\frac{\mu}{\log n}}} \\
&= \frac{1}{\mu^{\mu-\mu'}} \cdot \left(\frac{1}{1 - c_1\sqrt{\frac{1}{\mu\log n}}} \right)^{c_1\sqrt{\frac{\mu}{\log n}}} \\
&\leq \frac{1}{\mu^{\mu-\mu'}} \cdot \left(\frac{1}{\left(1-\frac{1}{2}\right)^2} \right)^{\frac{c_1^2}{\log n}} \\
&\leq \frac{2}{\mu^{\mu-\mu'}}.
\end{aligned}$$

□

Replacing the first term by the upper bound in the lemma implies

$$\frac{\mathrm{Var}(E_{x,\mu}^{\mu'})}{4} \leq \mathbb{E}\left[\left(\frac{(N_x')^{\underline{\mu-\mu'}}}{\mu^{\mu-\mu'}} \right)^2 \right] \mathbb{E}[\mathbb{1}_x^{\mu'}].$$

It suffices to bound the last quantity. To proceed, we need the following concentration inequalities for Poisson random variables . Note that these inequities hold for any Poisson random variables, and simply follow from the well-known Chernoff bound [MU05].

Lemma 6. *For $X \sim \mathrm{Poi}(M)$ and any $\lambda > 0$,*

$$\mathbb{P}(X \leq (1-\lambda)M) \leq \left(\frac{e^{-\lambda}}{(1-\lambda)^{(1-\lambda)}} \right)^M \leq e^{-\frac{\lambda^2 M}{2}},$$

and

$$\mathbb{P}(X \geq (1+\lambda)M) \leq \left(\frac{e^{\lambda}}{(1+\lambda)^{(1+\lambda)}}\right)^{M} \leq e^{-\frac{\min\{\lambda^2,\lambda\}M}{3}}.$$

Let c be a sufficiently large absolute constant. As a corollary of the lemma above, for any natural number $\mu > 100\log n$ and j such that $\sqrt{\mu/(c\log n)} > j \geq 1$,

$$\Pr\left(N'_x > \mu + jc\sqrt{\mu\log n}\right)\Pr(\mathbb{1}_x^{\mu} = 1) \leq e^{-\Theta(j\sqrt{c}\log n)},$$

and for any $i \geq 1$ and natural number μ,

$$\Pr\left(N'_x > \mu + i\mu\right)\Pr(\mathbb{1}_x^{\mu} = 1) \leq e^{-\Theta(i\mu)},$$

Intuitively, Poisson random variables are highly concentrated around their mean values. Hence, for a Poisson random variable X and natural numbers a, b such that $a \gg b$, we should expect the product $\Pr(X \geq a) \cdot \Pr(X \leq b)$ to be small. We are ready to bound the quantity of interest.

Lemma 7. *For sufficiently large n and any two natural numbers μ, μ' such that $n\log n > \mu > 100\log n$ and*

$$\mu - c_1\sqrt{\frac{\mu}{\log n}} \leq \mu' \leq \mu - 1,$$

we have

$$\mathbb{E}\left[\left(\frac{(N'_x)^{\underline{\mu-\mu'}}}{\mu^{\mu-\mu'}}\right)^2\right]\mathbb{E}[\mathbb{1}_x^{\mu'}] \leq e^{2c}\left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}\right).$$

Proof. The proof follows from the two concentration inequalities above. Note that for $\mu' \leq \mu - 1$,

those inequalities still hold if we replace $\Pr(\mathbb{1}_x^{\mu} = 1)$ by $\Pr(\mathbb{1}_x^{\mu'} = 1)$.

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{(N_x')^{\underline{\mu-\mu'}}}{\mu^{\mu-\mu'}}\right)^2\right]\mathbb{E}[\mathbb{1}_x^{\mu'}] \\
&\leq \left(1+c\sqrt{\frac{\log n}{\mu}}\right)^{2(\mu-\mu')}\mathbb{E}[\mathbb{1}_x^{\mu'}] + \sum_{j=1}^{\sqrt{\mu/(c\log n)}}\left(1+(j+1)c\sqrt{\frac{\log n}{\mu}}\right)^{2(\mu-\mu')} e^{-\Theta(j\sqrt{c}\log n)} \\
&+ \sum_{i=\sqrt{c}}^{\infty}(1+(i+1))^{2(\mu-\mu')}e^{-\Theta(i\mu)} \\
&\leq e^{2c}\mathbb{E}[\mathbb{1}_x^{\mu'}] + \sum_{j=1}^{\sqrt{\mu/(c\log n)}} e^{2(j+1)c}e^{-\Theta(j\sqrt{c}\log n)} + \sum_{i=\sqrt{c}}^{\infty} e^{\Theta(\sqrt{\mu}\log i)}e^{-\Theta(i\mu)} \\
&\leq e^{2c}\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}} + \sum_{i=\sqrt{c}}^{\infty} e^{\Theta(\sqrt{\mu}\log i)}e^{-\Theta(i\mu)} \\
&= e^{2c}\left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}\right).
\end{aligned}
$$

□

Ignoring the $1/n^{\Theta(\sqrt{c})}$ term, the proof actually shows that $E_{x,\mu}^{\mu'}$ is at most a constant multiple of $\sqrt{\mathbb{E}[\mathbb{1}_x^{\mu'}]}$, with high probability.

Under Poisson sampling, the multiplicity N_x is also a Poisson random variable with mean np_x. Note that $\mathbb{E}[\mathbb{1}_x^{\mu'}] = e^{-np_x}(np_x)^{\mu'}/\mu'! \leq (np_x)e^{-np_x}(np_x)^{\mu'-1}/(\mu'-1)! = \mathbb{E}[\mathbb{1}_x^{\mu'-1}](np_x)$. This observation together with an argument analogous to that above yields

Lemma 8. *Under the same conditions as in Lemma 4,*

$$\mathbb{E}\left[\left(\frac{(N_x')^{\underline{\mu-\mu'}}}{\mu^{\mu-\mu'}}\right)^2\right]\mathbb{E}[\mathbb{1}_x^{\mu'}] \leq e^{2c}\left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right).$$

Since $\mathbb{E}[\mathbb{1}_x^{\mu'}] = \Pr(N_x = \mu')$, and $E_{x,\mu}^{\mu'} = O(\sqrt{\mathbb{E}[\mathbb{1}_x^{\mu'}]})$ with high probability, there exists an absolute constant c' satisfying

$$\Pr(E_{x,\mu}^{\mu'} \geq c') \leq \frac{p_x}{n^{\Theta(c)}}.$$

An Estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$

While $E_{x,\mu}^{\mu'}$ is an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$, in the last section we showed that it can have a constant variance. To reduce the estimation variance, we estimate $\mathbb{E}[\mathbb{1}_x^\mu]$ by the following estimator

$$E_{x,\mu} := \frac{1}{c_1\sqrt{\mu/\log n}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E_{x,\mu}^{\mu'}.$$

The estimator simply averages a sequence of $E_{x,\mu}^{\mu'}$'s and remains as an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$. An important observation is that $E_{x,\mu}$ is the sum of $E_{x,\mu}^{\mu'} = \mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N_x')\overline{^{\mu-\mu'}}$, and only one of these terms can be non-zero, as $\mathbb{1}_x^{\mu'} \cdot \mathbb{1}_x^{\mu} = 0$ for all $\mu \neq \mu'$. Therefore, the inequality $\Pr(E_{x,\mu}^{\mu'} \geq c') \leq p_x/n^{\Theta(c)}$ immediately translates to

$$\Pr\left(E_{x,\mu} > \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \leq \frac{p_x}{n^{\Theta(c)}}.$$

We have designed $E_{x,\mu}$ in a way such that its variance would be small. Specifically,

Lemma 9. *Under the same conditions as in Lemma 4,*

$$\mathrm{Var}(E_{x,\mu}) \leq \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right).$$

Proof. The variance of $E_{x,\mu}$ satisfies

$$\begin{aligned}
\mathrm{Var}(E_{x,\mu}) &\overset{(a)}{=} \left(\frac{1}{c_1\sqrt{\mu/\log n}}\right)^2 \mathrm{Var}\left(\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N'_x)^{\underline{\mu-\mu'}}\right) \\
&\overset{(b)}{\le} \Theta\left(\frac{\log n}{\mu}\right) \mathbb{E}\left(\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N'_x)^{\underline{\mu-\mu'}}\right)^2 \\
&\overset{(c)}{=} \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{E}\left[\left(\mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N'_x)^{\underline{\mu-\mu'}}\right)^2\right] \\
&\overset{(d)}{\le} \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right),
\end{aligned}$$

where (a) follows from $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, (b) follows from $\mathrm{Var}(X) \le \mathbb{E}X^2$, (c) follows from $\mathbb{1}_x^{\mu'} \cdot \mathbb{1}_x^{\mu} = 0$ for all $\mu \neq \mu'$, and (d) follows from Lemma 1. □

Estimating $\mathbb{E}[\Phi_\mu]$

The last section shows that $E_{x,\mu}$ is a well-behaved estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$. Following the identity $\mathbb{E}[\Phi_\mu] = \sum_{x\in[k]} \mathbb{E}\left[\mathbb{1}_x^\mu\right]$, we naturally estimate $\mathbb{E}[\Phi_\mu]$ by

$$E_\mu := \sum_{x\in[k]} E_{x,\mu}.$$

By construction, E_μ is an unbiased estimator for $\mathbb{E}[\Phi_\mu]$. Due to Poisson sampling, all the

multiplicities N_x are independent. Following lemma 9, the variance of E_μ admits

$$\begin{aligned}\mathrm{Var}(E_\mu) &= \sum_{x\in[k]} \mathrm{Var}(E_{x,\mu}) \\ &\le \sum_{x\in[k]} \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right) \\ &= \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\Phi_{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}\right).\end{aligned}$$

Furthermore, by a non-asymptotic version of the Stirling's formula,

$$\mathbb{E}[\mathbb{1}_x^\mu] = e^{-np_x}\frac{(np_x)^\mu}{\mu!} \le e^{-\mu}\frac{\mu^\mu}{\mu!} \le e^{-\mu}\mu^\mu \cdot (2\pi)^{-1/2}\frac{e^\mu}{\mu^{\mu+1/2}} = \frac{1}{\sqrt{2\pi\mu}}.$$

Combining this with the following inequality mentioned in the last section,

$$\Pr\left(E_{x,\mu} > \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \le \frac{p_x}{n^{\Theta(c)}},$$

we immediately get

$$\Pr\left(\left|E_{x,\mu} - \mathbb{E}[\mathbb{1}_x^\mu]\right| > \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \le \Pr\left(E_{x,\mu} > -\frac{1}{\sqrt{2\pi\mu}} + \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \le \frac{p_x}{n^{\Theta(c)}},$$

where we have increased the value of c' by 1.

We are ready to characterize the tail probability of E_μ, for which we use the following variation [AJOS13b] of the well-known Bernstein inequality.

Lemma 10. *Let $Y_1,\ldots,Y_m$ be m independent variables such that with probability $\ge 1-\varepsilon_i$, $|Y_i - \mathbb{E}[Y_i]| < M$, then for any $\delta \in (0,1)$,*

$$\Pr\left(\left|\sum_i Y_i - \mathbb{E}\left[\sum_i Y_i\right]\right| > \sqrt{2\sum_i \mathrm{Var}(Y_i)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}\right) \le 2\delta + \sum_i \varepsilon_i.$$

Set $\delta = n^{-10}$, $m = k$, and $Y_x = E_{x,\mu}$ for all $x \in [k]$, and choose $M = c'/c_1\sqrt{\mu/\log n}$ and $\varepsilon_x = p_x/n^{\Theta(c)}$ for all $x \in [k]$. For a sufficiently large absolute constant c_4, the concentration inequality above combines all the previous results and yields

$$\Pr\left(\left|E_\mu - \mathbb{E}[\Phi_\mu]\right| > c_4 \frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}} \sqrt{\frac{1}{c_1^2} + \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{E}[\Phi_{\mu'}]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

Next we derive a similar inequality for which $\mathbb{E}[\Phi_{\mu'}]$'s in the inner sum are replaced with $\mathbb{E}[\Phi_\mu]$.

To do this, we utilize the following lemma [AJOS13b], which shows that $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu-1}]$ are often close to each other. Note that we have made the constants explicit.

Lemma 11. *For $\mu \geq 100\log n$,*

$$|\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu-1}]| \leq 5\sqrt{\frac{\log n}{\mu}}\mathbb{E}[\Phi_{\mu-1}] + \frac{3}{n^2},$$

and for $\mu \geq 1$,

$$\mathbb{E}[\Phi_\mu] \leq O\left((\log n)\mathbb{E}[\Phi_{\mu-1}] + \frac{1}{n}\right).$$

By the above lemma, for $n\log n > \mu \geq 100\log n$,

$$\mathbb{E}[\Phi_{\mu-1}] + \mu\frac{3}{n^2} \leq \left(1 + 10\sqrt{\frac{\log n}{\mu}}\right)\left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right).$$

This recursive inequality implies that for sufficiently small constant c_1 and any μ' satisfying

$\mu - c_1\sqrt{\mu/\log n} \le \mu' \le \mu - 1$,

$$\begin{aligned}
\mathbb{E}[\Phi_{\mu'}] + \mu'\frac{3}{n^2} &\le \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right)\prod_{i=\mu'+1}^{\mu}\left(1+10\sqrt{\frac{\log n}{i}}\right) \\
&\le \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right)\left(1+10\sqrt{\frac{\log n}{\mu - c_1\sqrt{\mu/\log n}}}\right)^{c_1\sqrt{\mu/\log n}} \\
&\le \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right)\left(1+\sqrt{\frac{121\log n}{\mu}}\right)^{11c_1\sqrt{\frac{\mu}{121\log n}}} \\
&\le \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right)e^{11c_1} \\
&\le 2\left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right),
\end{aligned}$$

where we have used the fact that $(1+1/x)^x < e$ for $x > 0$. Consequently, under the same conditions,

$$\mathbb{E}[\Phi_{\mu'}] \le 2\mathbb{E}[\Phi_\mu] + 2(\mu+1)\frac{3}{n^2} \le 2\mathbb{E}[\Phi_\mu] + \frac{7\log n}{n}.$$

Hence for sufficiently small c_1,

$$\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\mathbb{E}[\Phi_{\mu'}] \le 2c_1\sqrt{\mu/\log n}\left(\mathbb{E}[\Phi_\mu] + \frac{7\log n}{n}\right) \le 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu] + \frac{7\log n}{\sqrt{n}}.$$

This together with the previous tail bound yields

Lemma 12. *For $n\log n > \mu \ge 100\log n$,*

$$\Pr\left(\left|E_\mu - \mathbb{E}[\Phi_\mu]\right| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{1}{c_1^2} + 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right) \le \Theta\left(\frac{1}{n^{10}}\right).$$

An Alternative Estimator for $\mathbb{E}[\Phi_{\mu-1}]$

Under the proper conditions mentioned previously, $E_{x,\mu-1}$ is not only unbiased in estimating $\mathbb{E}[\mathbb{1}_x^{\mu-1}]$, but also has small variance. However, our latter analysis calls for bounding

the difference between $E_{x,\mu-1}$ and $E_{x,\mu}$, and it is inconvenient to use $E_{x,\mu-1}$ since it may have fewer terms than $E_{x,\mu}$. Hence to simplify our derivations, we construct the following estimator for $\mathbb{E}[\mathbb{1}_x^{\mu-1}]$,

$$E'_{x,\mu-1} := \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E^{\mu'}_{x,(\mu-1)},$$

and consequently estimate $\mathbb{E}[\Phi_{\mu-1}]$ by

$$E'_{\mu-1} := \sum_{x\in[k]} E'_{x,\mu-1}.$$

By an argument that is almost the same as that in the last few sections,

Lemma 13. *For $n\log n > \mu \geq 100\log n$,*

$$\Pr\left(\left|E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]\right| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{1}{c_1^2} + 2c_1\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_{\mu-1}]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

The Difference between Two Estimators

In this section, we consider

$$E^{(1)}_{\mu} := E_{\mu} - E'_{\mu-1} = \sum_{x\in[k]} (E_{x,\mu} - E'_{x,\mu-1}),$$

the difference between the two estimators E_μ and $E'_{\mu-1}$. We show that, $E^{(1)}_{\mu}$, as an unbiased estimator for $\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu-1}]$, highly concentrates around its mean. In the subsequent sections, we leverage this property to design an accurate estimator for the total probability M_μ.

Similar to the previous derivations, we start by considering a single term

$$E^{(1)}_{x,\mu} := E_{x,\mu} - E'_{x,\mu-1}.$$

We can bound the absolute value of $E^{(1)}_{x,\mu}$ as follows.

$$\begin{aligned}
|E^{(1)}_{x,\mu}| &= |E_{x,\mu} - E'_{x,\mu-1}| \\
&= \left| \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E^{\mu'}_{x,\mu} - \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E^{\mu'}_{x,(\mu-1)} \right| \\
&\le \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left| E^{\mu'}_{x,\mu} - E^{\mu-1}_{x,\mu} \right| \\
&= \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left| \mathbb{1}^{\mu'}_x a^{\mu'}_\mu (N'_x)^{\underline{\mu-\mu'}} - \mathbb{1}^{\mu'}_x a^{\mu'}_{\mu-1} (N'_x)^{\underline{(\mu-1)-\mu'}} \right| \\
&= \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{1}^{\mu'}_x a^{\mu'}_\mu (N'_x)^{\underline{(\mu-1)-\mu'}} \left| (N'_x - \mu) - (\mu - \mu') + 1 \right|.
\end{aligned}$$

The above inequality together with $\mathrm{Var}(E^{(1)}_{x,\mu}) \le \mathbb{E}(E^{(1)}_{x,\mu})^2$ implies

$$\mathrm{Var}(E^{(1)}_{x,\mu}) \le \frac{1}{c_1^2}\frac{\log n}{\mu} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} (a^{\mu'}_\mu)^2 \mathbb{E}[\mathbb{1}^{\mu'}_x] \mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}} \left| (N'_x - \mu) - (\mu - \mu') + 1 \right| \Big)^2,$$

where we have used $\mathbb{1}^{\mu'}_x \cdot \mathbb{1}^{\mu}_x = 0$ for all $\mu \neq \mu'$. Note that the bound on the right-hand side is a sum of three-term products. Assume that $n \gg 1$ and $n\log n > \mu > 100\log n$, and consider one of these products that corresponds to an arbitrary μ' satisfying $\mu - c_1(\mu/\log n) \le \mu' \le \mu - 1$. Lemma 5 bounds its first term as $(a^{\mu'}_\mu)^2 \le 4/(\mu^{\mu-\mu'})^2$. Replacing the first term with this bound, the following lemma further upper bounds the resulting quantity.

Lemma 14. *Under the same conditions as in Lemma 4,*

$$\frac{\mathbb{E}[\mathbb{1}^{\mu'}_x]}{(\mu^{\mu-\mu'})^2} \mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}} \left| (N'_x - \mu) - (\mu - \mu') + 1 \right| \Big)^2 \le \Theta\left(\frac{\log n}{\mu}\right) \mathbb{E}[\mathbb{1}^{\mu'}_x] + \frac{1}{n^{\Theta(\sqrt{c})}}.$$

Proof. Since $(N'_x - \mu)$ can be negative, we need the concentration inequality

$$\Pr\left(N'_x - \mu < -c\sqrt{\mu \log n}\right) \Pr(\mathbb{1}_x^\mu = 1) \le e^{-\Theta(\sqrt{c}\log n)},$$

which follows from Lemma 6. Similar to the proof of Lemma 4, we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_x^{\mu'}]\mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}}\big|(N'_x-\mu)-(\mu-\mu')+1\big|\Big)^2\frac{1}{(\mu^{\mu-\mu'})^2}\\
&\le \mathbb{E}[\mathbb{1}_x^{\mu'}]\left(\frac{2c\sqrt{\mu\log n}}{\mu}\right)^2\left(1+c\sqrt{\frac{\log n}{\mu}}\right)^{2((\mu-1)-\mu')}+\left(\frac{2\mu-1}{u}\right)^2 e^{-\Theta(\sqrt{c}\log n)}\\
&+\sum_{j=1}^{\sqrt{\mu/(c\log n)}}\left(1+(j+1)c\sqrt{\frac{\log n}{\mu}}\right)^{2(\mu-\mu')}e^{-\Theta(j\sqrt{c}\log n)}+\sum_{i=\sqrt{c}}^{\infty}(1+(i+1))^{2(\mu-\mu')}e^{-\Theta(i\mu)}.
\end{aligned}$$

Since $(1+1/x)^x < e$ for $x > 0$, the first term on the right-hand side can be bounded by $4(c^2 \cdot e^{2c})\mathbb{E}[\mathbb{1}_x^{\mu'}](\log n)/\mu$. The sum of the remaining three terms is at most

$$4e^{-\Theta(\sqrt{c}\log n)}+\sum_{j=1}^{\sqrt{\mu/(c\log n)}}e^{2(j+1)c}e^{-\Theta(j\sqrt{c}\log n)}+\sum_{i=\sqrt{c}}^{\infty}e^{\Theta(\sqrt{\mu}\log i)}e^{-\Theta(i\mu)}\le\frac{1}{n^{\Theta(\sqrt{c})}}.$$

Consolidating these bounds yields the desired result. □

By $\mathbb{E}[\mathbb{1}_x^{\mu'}] \le \mathbb{E}[\mathbb{1}_x^{\mu'-1}](np_x)$, an analogous argument yields

Lemma 15. *Under the same conditions as in Lemma 4,*

$$\frac{\mathbb{E}[\mathbb{1}_x^{\mu'}]}{(\mu^{\mu-\mu'})^2}\mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}}\big|(N'_x-\mu)-(\mu-\mu')+1\big|\Big)^2\le\Theta\left(\frac{\log n}{\mu}\right)\mathbb{E}[\mathbb{1}_x^{\mu'}]+\frac{p_x}{n^{\Theta(\sqrt{c})}}.$$

There is always a unique μ' such that $\mathbb{1}_x^{\mu'} = 1$. The proof of Lemma 14 together with $\mathbb{E}[\mathbb{1}_x^{\mu'}] \le \mathbb{E}[\mathbb{1}_x^{\mu'-1}](np_x)$ also shows that for a sufficiently large absolute constant c'',

$$\Pr\left(|E_{x,\mu}^{(1)}| > \frac{c''\log n}{\mu}\right) \le \frac{p_x}{n^{\Theta(c)}}.$$

Furthermore, the expectation of $E^{(1)}_{x,\mu}$ satisfies

Lemma 16. *For any natural number μ such that $n\log n > \mu \geq 100\log n$,*

$$|\mathbb{E}[E^{(1)}_{x,\mu}]| \leq 1/\mu.$$

Proof. Recall that $E^{(1)}_{x,\mu}$ is an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^{\mu}] - \mathbb{E}[\mathbb{1}_x^{\mu-1}]$. Therefore,

$$\begin{aligned}
|\mathbb{E}[E^{(1)}_{x,\mu}]| &= |\mathbb{E}[\mathbb{1}_x^{\mu}] - \mathbb{E}[\mathbb{1}_x^{\mu-1}]| \\
&= \left| e^{-np_x}\frac{(np_x)^{\mu}}{\mu!} - e^{-np_x}\frac{(np_x)^{\mu-1}}{(\mu-1)!} \right| \\
&= \left| e^{-np_x}\frac{(np_x)^{\mu-1}(np_x-\mu)}{\mu!} \right|.
\end{aligned}$$

In general, consider the function $g_\mu(y) := e^{-y}y^{\mu-1}(y-\mu)/\mu!$ for $y \geq 0$. The first-order derivative of $g_\mu(y)$ with respect to y is

$$g'_\mu(y) = -\frac{1}{\mu!}e^{-y}y^{-2+\mu}(\mu^2 + y^2 - \mu(1+2y))$$

which has two roots, $y_1 := \mu - \sqrt{\mu}$ and $y_2 := \mu + \sqrt{\mu}$. Since both $g_\mu(0)$ and $\lim_{y\to\infty} g(y)$ equal to zero, the maximum of $|g_\mu(y)|$ for $y \geq 0$ is $\max\{|g(y_1)|, |g(y_2)|\}$. By a non-asymptotic version of

the Stirling's formula,

$$\begin{aligned}
|g(y_1)| &= e^{-\mu+\sqrt{\mu}}\frac{(\mu-\sqrt{\mu})^{\mu-1}\sqrt{\mu}}{\mu!} \\
&\le e^{-\mu+\sqrt{\mu}}(\mu-\sqrt{\mu})^{\mu-1}\sqrt{\mu}\frac{e^{\mu}}{\sqrt{2\pi}\mu^{\mu+\frac{1}{2}}} \\
&= e^{\sqrt{\mu}}\left(\frac{\mu-\sqrt{\mu}}{\mu}\right)^{\mu-1}\frac{1}{\sqrt{2\pi}\mu} \\
&= e^{\sqrt{\mu}}\left(1-\frac{1}{\sqrt{\mu}}\right)^{\sqrt{\mu}(\sqrt{\mu}-(1/\sqrt{\mu}))}\frac{1}{\sqrt{2\pi}\mu} \\
&\le e^{\sqrt{\mu}}e^{-(\sqrt{\mu}-(1/\sqrt{\mu}))}\frac{1}{\sqrt{2\pi}\mu} \\
&= \frac{e^{1/\sqrt{\mu}}}{\sqrt{2\pi}\mu} \le \frac{1}{\mu}.
\end{aligned}$$

Similarly, we can also show that $|g(y_1)| \le 1/\mu$. □

Increase the value of c'' by 1. The above lemma implies

$$\Pr\left(|E^{(1)}_{x,\mu} - \mathbb{E}[E^{(1)}_{x,\mu}]| > \frac{c''\log n}{\mu}\right) \le \frac{p_x}{n^{\Theta(c)}}.$$

Turning back to $E^{(1)}_{\mu}$ and using Lemma 15, we can bound the variance of $E^{(1)}_{\mu}$ as

$$\mathrm{Var}(E^{(1)}_{\mu}) \le \sum_{x\in[k]}\mathrm{Var}(E^{(1)}_{x,\mu}) \le \Theta\left(\frac{\log^2 n}{\mu^2}\right)\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\left(\mathbb{E}[\Phi_{\mu'}] + \frac{1}{n^{\Theta((c/k)\wedge k)}}\right).$$

Let c'_4 be a sufficiently large absolute constant. By the Bernstein-inequality variation in Lemma 10,

$$\Pr\left(\left|E^{(1)}_{\mu} - \mathbb{E}[E^{(1)}_{\mu}]\right| > c'_4\frac{\log^2 n}{\mu}\sqrt{1+\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\mathbb{E}[\Phi_{\mu'}]}\right) \le \Theta\left(\frac{1}{n^{10}}\right).$$

By Lemma 11, for sufficiently small constant c_1 and any μ' satisfying $\mu - c_1\sqrt{\mu/\log n} \le \mu' \le \mu-1$

,

$$\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\mathbb{E}[\Phi_{\mu'}] \le 2c_1\sqrt{\mu/\log n}\left(\mathbb{E}[\Phi_\mu]+\frac{7\log n}{n}\right) \le 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]+\frac{7\log n}{\sqrt{n}}.$$

Combined, the two inequalities above yield

Lemma 17. *For $n\log n > \mu \ge 100\log n$,*

$$\Pr\left(\left|E_\mu^{(1)}-\mathbb{E}[E_\mu^{(1)}]\right| > c_4'\frac{\log^2 n}{\mu}\sqrt{2+2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right) \le \Theta\left(\frac{1}{n^{10}}\right).$$

Estimating the Total Probability Mass

A genie estimator that knows both $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu-1}]$ could accurately estimate $M_{\mu-1}$ by

$$O_{\mu-1} := \Phi_{\mu-1}\frac{\mu}{n}\frac{\mathbb{E}[\Phi_\mu]}{\mathbb{E}[\Phi_{\mu-1}]}.$$

and achieve the following guarantee [AJOS13b] for a sufficiently large constant c_4''.

Lemma 18. *For μ satisfying $n\log n > \mu \ge 100\log n$ and $\mathbb{E}[\Phi_{\mu-1}] \ge 1$,*

$$\Pr\left(|M_{\mu-1}-O_{\mu-1}| \ge c_4''\frac{\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n}{n}\right) \le O\left(\frac{1}{n^{10}}\right).$$

Replace $\mathbb{E}[\Phi_\mu]/\mathbb{E}[\Phi_{\mu-1}]$ with $E_\mu/E'_{\mu-1}$. Our estimator is simply

$$\hat{O}_{\mu-1} := \Phi_{\mu-1}\frac{\mu}{n}\frac{E_\mu}{E'_{\mu-1}}.$$

Note that we use $E'_{\mu-1}$ instead of $E_{\mu-1}$ just to simplify the proofs. Clearly, our objective is to characterize the estimation error $|M_{\mu-1}-\hat{O}_{\mu-1}|$. By the triangle inequality and the above lemma, it suffices to bound $\left|O_{\mu-1}-\hat{O}_{\mu-1}\right|$. To do this, we use the following interesting result.

Lemma 19. *If $b > 0$, $b + \Delta b > 0$, and $|\Delta b| \leq 0.9b$,*

$$\left| \frac{a+\Delta a}{b+\Delta b} - \frac{a}{b} \right| \leq O\left(\frac{|\Delta b||a| + |\Delta a||b|}{b^2} \right).$$

The above lemma appears in [AJOS13b] and follows by simple algebra. Set $a = \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, $b = \mathbb{E}[\Phi_{\mu-1}]$, $\Delta a = E_\mu - E'_{\mu-1} - \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, and $\Delta b = E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]$. Note that $a = \mathbb{E}[E_\mu^{(1)}]$ and $\Delta a = E_\mu^{(1)} - \mathbb{E}[E_\mu^{(1)}]$. Assuming that $n \log n > \mu \geq 100 \log n$, we analyze each term below.

For $a = \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, by Lemma 11,

$$|a| \leq 5\sqrt{\frac{\log n}{\mu}}\mathbb{E}[\Phi_{\mu-1}] + \frac{3}{n^2}.$$

For $\Delta a = E_\mu^{(1)} - \mathbb{E}[E_\mu^{(1)}]$, as shown in Lemma 17,

$$\Pr\left(|\Delta a| > c_4' \frac{\log^2 n}{\mu} \sqrt{2 + 2c_1 \sqrt{\frac{\mu}{\log n}} \mathbb{E}[\Phi_\mu]} \right) \leq \Theta\left(\frac{1}{n^{10}} \right).$$

For $b = \mathbb{E}[\Phi_{\mu-1}]$, Lemma 11 implies a lower bound

$$b \geq \left(1 + 5\sqrt{\frac{\log n}{\mu}} \right)^{-1} \left(\mathbb{E}[\Phi_\mu] - \frac{3}{n^2} \right) \geq \frac{2}{3}\mathbb{E}[\Phi_\mu] - \frac{2}{n^2},$$

as well as an upper bound

$$b \leq \left(1 + 10\sqrt{\frac{\log n}{\mu}} \right) \mathbb{E}[\Phi_\mu] + \frac{3}{n^2} \leq 2\mathbb{E}[\Phi_\mu] + \frac{3}{n^2}.$$

For $\Delta b = E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]$, Lemma 13 states that

$$\Pr\left(|\Delta b| > c_4 \frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}} \sqrt{\frac{1}{c_1^2} + 2c_1\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_{\mu-1}]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

Our bound on $|b|$ further implies

$$\Pr\left(|\Delta b| > c_4 \frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}} \sqrt{\frac{2}{c_1^2} + 6c_1\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_{\mu}]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

Here, we can choose a sufficiently large constant c_5 so that, if $n \gg 1$, $\mu > 100\log n$, and $\mathbb{E}[\Phi_{\mu-1}] > c_5(\log^2 n)/10$, then $|\Delta b| < 0.9b$ with probability at least $1 - \Theta(n^{-10})$. Also note that $\Phi_\mu = \sum_{x\in[k]} \mathbb{1}_x^\mu$. In Lemma 10, set $\delta = n^{-10}$, $m = k$, and $Y_x = \mathbb{1}_x^\mu$ for all $x \in [k]$, $M = 1$, and choose $\varepsilon_x = 0$ for all $x \in [k]$. Then,

$$\Pr\left(\left|\Phi_\mu - \mathbb{E}\left[\Phi_\mu\right]\right| > \sqrt{20\sum_{x\in[k]} \mathrm{Var}(\mathbb{1}_x^\mu)\log n} + \frac{20}{3}\log n\right) \leq \frac{2}{n^{10}}.$$

Together with $\mathrm{Var}(\mathbb{1}_x^\mu) \leq \mathbb{E}(\mathbb{1}_x^\mu)^2 = \mathbb{E}[\mathbb{1}_x^\mu]$, the above inequality implies

Lemma 20. *For sufficiently large n, and μ satisfying $\mu > 100\log n$ and $\mathbb{E}[\Phi_{\mu-1}] > c_5(\log^2 n)/10$,*

$$\Pr\left(\left|\Phi_\mu - \mathbb{E}\left[\Phi_\mu\right]\right| > \sqrt{20\mathbb{E}[\Phi_\mu]\log n} + \frac{20}{3}\log n\right) \leq \frac{2}{n^{10}}.$$

For our purpose, it suffices to apply the estimator $\hat{O}_{\mu-1}$ to indices μ satisfying $\mu > 100\log n$ and $\mathbb{E}[\Phi_\mu] \geq 0.5c_5\log^2 n$. While not knowing p, we can use the independent sample sequence $X^{N''}$ to ensure that with high probability, $\mathbb{E}[\Phi_\mu] \geq 0.5c_5\log^2 n$. More concretely, we only apply $\hat{O}_{\mu-1}$ to indices μ satisfying $\Phi''_{\mu-1} > c_5\log^2 n$. By construction, $\mathbb{E}\left[\Phi_\mu\right] = \mathbb{E}\left[\Phi''_\mu\right]$. Then for

sufficiently large c_5 and n, and $\mathbb{E}[\Phi_{\mu-1}] < 0.5c_5 \log^2 n$, Lemma 20 implies

$$\Pr\left(\Phi''_{\mu-1} > c_5 \log^2 n\right) \leq \Pr\left(\left|\Phi''_{\mu-1} - \mathbb{E}\left[\Phi_{\mu-1}\right]\right| > \sqrt{20\mathbb{E}[\Phi_{\mu-1}]\log n} + \frac{20}{3}\log n\right) \leq \frac{2}{n^{10}}.$$

Hence for μ satisfying the conditions mentioned previously, we can assume that $\mathbb{E}[\Phi_{\mu-1}] \geq 0.5c_5 \log^2 n$. Under this assumption, Lemma 11 implies that $\mathbb{E}[\Phi_\mu] \geq \mathbb{E}[\Phi_{\mu-1}]/3 \geq c_5(\log^2 n)/6$. By the same reasoning, $\mathbb{E}[\Phi_\mu]/18 \leq \mathbb{E}[\Phi_{\mu-1}]/6 \leq \Phi_{\mu-1} \leq 6\mathbb{E}[\Phi_{\mu-1}] \leq 18\mathbb{E}[\Phi_\mu]$ with probability at least $1 - \Theta(n^{-10})$. In other words, we can also assume that $\Phi_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]) = \Theta(\mathbb{E}[\Phi_\mu])$.

Recall that $a = \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, $b = \mathbb{E}[\Phi_{\mu-1}]$, $\Delta a = E_\mu - E'_{\mu-1} - \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, and $\Delta b = E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]$. The union bound together with Lemma 19 combines all the results in this section and yields that with probability at least $1 - \Theta(n^{-10})$,

$$\begin{aligned}
\left|O_{\mu-1} - \hat{O}_{\mu-1}\right| &= \Phi_{\mu-1}\frac{\mu}{n}\left|\frac{E_\mu}{E_{\mu-1}} - \frac{\mathbb{E}[\Phi_\mu]}{\mathbb{E}[\Phi_{\mu-1}]}\right| \\
&= \Phi_{\mu-1}\frac{\mu}{n}\left|\frac{E_\mu - E_{\mu-1}}{E_{\mu-1}} - \frac{\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu-1}]}{\mathbb{E}[\Phi_{\mu-1}]}\right| \\
&= \Phi_{\mu-1}\frac{\mu}{n}\left|\frac{a+\Delta a}{b+\Delta b} - \frac{a}{b}\right| \\
&\leq O\left(\Phi_{\mu-1}\frac{\mu}{n}\frac{|\Delta b||a| + |\Delta a||b|}{b^2}\right) \\
&\leq O\left(\Phi_{\mu-1}\frac{\mu}{n}\frac{\left(\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_\mu]}\right)\sqrt{\frac{\log n}{\mu+1}}\mathbb{E}[\Phi_{\mu-1}]}{(\mathbb{E}[\Phi_\mu])^2}\right) \\
&\quad + O\left(\Phi_{\mu-1}\frac{\mu}{n}\frac{\left(\frac{\log^2 n}{\mu}\sqrt{\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right)\mathbb{E}[\Phi_\mu]}{(\mathbb{E}[\Phi_\mu])^2}\right) \\
&= O\left(\frac{\log^2 n}{n}\sqrt{\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right).
\end{aligned}$$

Again, we can make c_5 sufficiently large so that with probability at least $1 - O(n^{-10})$, the upper

bound is at most $0.9O_{\mu-1}$ and $O_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]\mu/n)$. Combined, the upper bound of $0.9O_{\mu-1}$, the identity $\Phi_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]) = \Theta(\mathbb{E}[\Phi_\mu])$, and Lemma 18 imply

Lemma 21. *For μ satisfying $n\log n > \mu \geq 100\log n$ and $\Phi''_\mu > c_5\log^2 n$,*

$$\Pr\left(|M_{\mu-1} - \hat{O}_{\mu-1}| \geq 2c''_4\frac{\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n}{n}\right) \leq O\left(\frac{1}{n^{10}}\right).$$

Therefore, with probability at least $1 - O(n^{-10})$, we have both $\hat{O}_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]\mu/n)$ and

$$|M_{\mu-1} - \hat{O}_{\mu-1}| \leq O\left(\frac{\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n}{n}\right) = O\left(\frac{\sqrt{\Phi_{\mu-1}(\mu-1)}\log^2 n}{n}\right).$$

Furthermore, if these two claims hold,

$$\frac{(M_{\mu-1} - \hat{O}_{\mu-1})^2}{\hat{O}_{\mu-1}} \leq O\left(\frac{\left((\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n)/n\right)^2}{\mathbb{E}[\Phi_{\mu-1}]\mu/n}\right) \leq O\left(\frac{\log^4 n}{n}\mathbb{1}_{\Phi_{\mu-1}>0}\right).$$

Finally, we note that these results hold with high probability, i.e., $1 - O(n^{-10})$, instead of surely. To make sure that the KL-divergence between the underlying truth and our estimates is not infinity, we modify our estimator slightly and denote

$$\hat{O}'_{\mu-1} := \min\{\max\{1/n, \hat{O}_{\mu-1}\}, \log^2 n\}.$$

We use $\hat{O}'_{\mu-1}$ to estimate $M_{\mu-1}$ iff μ satisfies $n\log n > \mu \geq 100\log n$ and $\Phi''_\mu > 2c_5\log^2 n$. Note that this estimator also admits the above inequalities, since with probability at least $1 - O(n^{-10})$, the value of the original estimator satisfies $\hat{O}_{\mu-1} = \Theta(\mu\Phi_{\mu-1}/n) \leq O(N/n) = O(\log n) < \log^2 n$ and $\hat{O}_{\mu-1} = \Theta(\mu\Phi_{\mu-1}/n) \geq \Omega((\log^3 n)/n) > 1/n$, implying that $\hat{O}'_{\mu-1} = \hat{O}_{\mu-1}$.

The Good-Turing Estimator

The Good-Turing estimator estimates $M_{\mu-1}$ by

$$\hat{G}_{\mu-1} := \frac{\mu}{n}\Phi_{\mu}.$$

Let c_5' be a sufficiently large absolute constant. The following lemma [DM04] characterizes the performance of $\hat{G}_{\mu-1}$ in estimating $M_{\mu-1}$.

Lemma 22. *For μ satisfying $\mathbb{E}[\Phi_{\mu-1}] \geq 1$ and $\delta \in (0,1)$,*

$$\Pr\left(|M_{\mu-1} - \hat{G}_{\mu-1}| > c_5'\sqrt{\mathbb{E}[\Phi_{\mu}]+1}\frac{\mu\log^2\frac{n}{\delta}}{n}\right) \leq \delta.$$

For indices μ satisfying $2 \leq u \leq 100\log n$ and $\Phi''_{\mu-1} > 2c_5(\log^2 n)$, we simply use the following variant of the Good-Turing estimator,

$$\hat{G}'_{\mu-1} := \max\left\{\frac{1}{n}, \hat{G}_{\mu-1}\right\}.$$

Given $\Phi''_{\mu-1} > 2c_5(\log^2 n)$, by derivations in the last section, we can assume that $\Phi_{\mu-1} = \Theta(\Phi_{\mu}) = \Theta(\mathbb{E}[\Phi_{\mu-1}]) = \Theta(\mathbb{E}[\Phi_{\mu}]) \geq \log^2 n$, and with probability at least $1 - O(n^{-10})$, we would be correct. Choose $\delta = n^{-10}$ in Lemma 22. Then,

$$\Pr\left(|M_{\mu-1} - \hat{G}_{\mu-1}| > 15^2 c_5'\sqrt{\mathbb{E}[\Phi_{\mu}]}\frac{\mu\log^2 n}{n}\right) \leq \frac{1}{n^{10}}.$$

Additionally, note that $\mu \leq 100\log n$. Hence with probability at least $1 - O(n^{-10})$,

$$|M_{\mu-1} - \hat{G}_{\mu-1}| \leq O\left(\sqrt{\mathbb{E}[\Phi_{\mu}]}\frac{\mu\log^2 n}{n}\right) \leq O\left(\frac{\sqrt{\Phi_{\mu-1}(\mu-1)}\log^{5/2} n}{n}\right),$$

and

$$\frac{(M_{\mu-1}-\hat{G}_{\mu-1})^2}{\hat{G}_{\mu-1}} \le O\left(\frac{\left(\sqrt{\mathbb{E}[\Phi_\mu]}\frac{\mu\log^2 n}{n}\right)^2}{\frac{\mu}{n}\mathbb{E}[\Phi_\mu]}\right) = O\left(\frac{\mu\log^4 n}{n}\right) = O\left(\frac{\log^5 n}{n}\mathbb{1}_{\Phi_{\mu-1}>0}\right).$$

The estimator $G'_{\mu-1}$ also admits these inequalities since with probability at least $1-O(n^{-10})$, we have $G'_{\mu-1} = \mu\Phi_\mu/n \ge 2(\log^3 n)/n > 1/n$, implying $G'_\mu = G_\mu$.

An Estimator for M_0

For $\mu = 1$, regardless of the value of $\Phi''_{\mu-1}$, we estimate the total probability $M_{\mu-1} = M_0$, by the estimator $\hat{G}'_0 = \max\{1,\Phi_1\}/n$. We divide our analysis into two cases according to $\mathbb{E}[\Phi_0]$.

Case 1: If $\mathbb{E}[\Phi_0] \ge 1$, then by Lemma 22, with probability at least $1-O(n^{-10})$,

$$|M_0-\hat{G}'_0| \le |M_0-\hat{G}_0| + \frac{1}{n} \le O\left(\sqrt{\mathbb{E}[\Phi_1]+1}\frac{\log^2 n}{n}\right)$$

If $\mathbb{E}[\Phi_1] \ge c_5\log^2 n$, then by Lemma 20 and arguments in the last section, with probability at least $1-O(n^{-10})$, we have $\mathbb{E}[\Phi_1] = \Theta(\Phi_1) \ge \Omega(\log^2 n)$. This together with the above inequality further implies $\hat{G}'_0 = \Phi_1/n$ and

$$|M_0-\hat{G}'_0| \le O\left(\sqrt{\Phi_1}\frac{\log^2 n}{n}\right).$$

Therefore, with probability at least $1-O(n^{-10})$, we have $\Phi_1 > 0$ and

$$\frac{(M_0-\hat{G}'_0)^2}{\hat{G}'_0} \le O\left(\frac{\left(\sqrt{\Phi_1}\frac{\log^2 n}{n}\right)^2}{\Phi_1/n}\right) \le O\left(\frac{\log^4 n}{n}\mathbb{1}_{\Phi_1>0}\right).$$

If $\mathbb{E}[\Phi_1] < c_5 \log^2 n$, then by the first inequality, with probability at least $1 - O(n^{-10})$,

$$|M_0 - \hat{G}_0'| \leq O\left(\sqrt{\mathbb{E}[\Phi_1] + 1}\frac{\log^2 n}{n}\right) \leq O\left(\frac{\log^3 n}{n}\right),$$

which further implies

$$\frac{(M_0 - \hat{G}_0')^2}{\hat{G}_0'} \leq O\left(\frac{\left(\frac{\log^3 n}{n}\right)^2}{1/n}\right) \leq O\left(\frac{\log^6 n}{n}\right).$$

Case 2: If $\mathbb{E}[\Phi_0] \leq 1$, then by Lemma 11,

$$\mathbb{E}[\Phi_1] \leq O\left((\log n)\mathbb{E}[\Phi_0] + \frac{1}{n}\right) \leq O(\log n).$$

Furthermore, by Lemma 20, with probability at least $1 - O(n^{-10})$,

$$\Phi_0 \leq O(\log n).$$

For $\delta \in (0,1)$ and symbols x satisfying $p_x \geq \log(n/\delta)/n$, we have $\Pr(\mathbb{1}_x^0 = 1) = e^{-np_x} \leq \delta/n$. Note that the number of such symbols is at most n. Hence by the union bound,

$$\Pr\left(\exists x \in [k] \text{ s.t. } p_x > \frac{\log(n/\delta)}{n}, \mathbb{1}_x^0 = 1\right) \leq n \cdot \frac{\delta}{n} = \delta.$$

Setting $\delta = n^{-10}$ in the above inequality yields

$$\Pr\left(\forall x \in [k] \text{ s.t. } p_x > \frac{11\log(n)}{n}, \mathbb{1}_x^0 = 0\right) \geq 1 - n^{-10}.$$

Therefore if we further have $\Phi_0 \leq O(\log n)$,

$$M_0 = \sum_{x\in[k]} \mathbb{1}_x^0 \cdot p_x \leq O(\Phi_0) \cdot \frac{11\log(n)}{n} = O\left(\frac{\log^2 n}{n}\right).$$

In addition, since $\mathbb{E}[\Phi_1] \leq O(\log n)$, Lemma 20 implies that with probability at least $1 - O(n^{-10})$,

$$\Phi_1 \leq O(\log n).$$

Consolidating these results shows that with probability at least $1 - O(n^{-10})$,

$$|M_0 - \hat{G}_0'| \leq O\left(\frac{\log^2 n}{n} + \frac{\log n}{n}\right) = O\left(\frac{\log^2 n}{n}\right).$$

and

$$\frac{(M_0 - \hat{G}_0')^2}{\hat{G}_0'} \leq O\left(\frac{\left(\frac{\log^2 n}{n}\right)^2}{1/n}\right) \leq O\left(\frac{\log^4 n}{n}\right).$$

Summary of case 1 and 2: With probability at least $1 - O(n^{-10})$,

$$|M_0 - \hat{G}_0'| = O\left(\frac{(\sqrt{\Phi_1}+1)\log^3 n}{n}\right)$$

and

$$\frac{(M_0 - \hat{G}_0')^2}{\hat{G}_0'} \leq O\left(\frac{\log^6 n}{n}\right).$$

The Empirical Estimator

For $\Phi''_{\mu-1} \leq 2c_5(\log^2 n)$ and $\mu \geq 2$, we use the empirical estimator,

$$\hat{\phi}_{\mu-1} := \frac{\mu-1}{n}\Phi_{\mu-1}.$$

By Lemma 20, since $\Phi''_{\mu-1} \leq 2c_5(\log^2 n)$, we can assume that $\mathbb{E}[\Phi_{\mu-1}] \leq O(\log^2 n)$ and $\Phi_{\mu-1} \leq O(\log^2 n)$, and be correct with probability at least $1 - O(n^{-10})$.

The next lemma in [DM04] characterizes the performance of $\hat{\phi}_{\mu-1}$ in estimating $M_{\mu-1}$.

Lemma 23. *For $\mu \geq 2$ and $\delta \in (0,1)$,*

$$\Pr\left(|M_{\mu-1} - \hat{\phi}_{\mu-1}| \leq O\left(\Phi_{\mu-1}\frac{\sqrt{\mu}\log\frac{n}{\delta}}{n}\right)\right) \geq 1-\delta.$$

Setting $\delta = n^{-10}$ in the lemma implies that with probability at least $1 - n^{-10}$,

$$|M_{\mu-1} - \hat{\phi}_{\mu-1}| \leq O\left(\Phi_{\mu-1}\frac{\sqrt{\mu}\log n}{n}\right) \leq O\left(\frac{\sqrt{\Phi_{\mu-1}(\mu-1)}\log^2 n}{n}\right).$$

Assume that all the inequalities above hold. Then,

$$\frac{(M_{\mu-1} - \hat{\phi}_{\mu-1})^2}{\hat{\phi}_{\mu-1}} \leq O\left(\frac{\left(\Phi_{\mu-1}\frac{\sqrt{\mu}\log n}{n}\right)^2}{\frac{\mu-1}{n}\Phi_{\mu-1}}\right) = O\left(\frac{\Phi_{\mu-1}\log^2 n}{n}\right) = O\left(\frac{\log^4 n}{n}\mathbb{1}_{\Phi_{\mu-1}>0}\right).$$

As a final remark, we can choose $c_2 = 2c_5$.

Final Estimator

In case our estimates sum to 1, we can simply estimate each M_μ by

$$\hat{M}_\mu := \begin{cases} \hat{G}'_\mu & \text{if } \mu = 0, \\ \hat{\phi}_\mu & \text{if } \mu \geq 1 \text{ and } \Phi_\mu \leq c_2(\log^2 n), \\ \hat{O}'_\mu & \text{if } \mu > c_3 \log n \text{ and } \Phi_\mu > c_2(\log^2 n), \\ \hat{G}'_\mu & \text{if } c_3 \log n \geq \mu \geq 1 \text{ and } \Phi_\mu > c_2(\log^2 n), \end{cases}$$

Otherwise, we normalize these probability estimates by their sum,

$$T := \sum_{\mu \geq 0} \hat{M}_\mu,$$

and approximate each M_μ by $\hat{M}^*_\mu := \hat{M}_\mu / T$.

First we show that T is often close to 1. By Lemma 6, under Poisson sampling,

$$\Pr\left(1 \leq \sum_{\mu \geq 1} \Phi_\mu \mu = \mathrm{Poi}(n) \leq n \log n\right) \geq 1 - O(e^{-n}).$$

By the union bound and results in the previous sections, with probability at least $1 - O(n^{-8})$,

$$|M_\mu - \hat{M}_\mu| \leq \tilde{O}\left(\frac{\sqrt{\Phi_\mu \mu}}{n}\right), \forall \mu \geq 1,$$

$$\frac{(M_\mu - \hat{M}_\mu)^2}{\hat{M}_\mu} \leq \tilde{O}\left(\frac{\mathbb{1}_{\Phi_\mu > 0}}{n}\right), \forall \mu \geq 1,$$

$$|M_0 - \hat{M}_0| \leq \tilde{O}\left(\frac{\sqrt{\Phi_1} + 1}{n}\right),$$

and

$$\frac{(M_0 - \hat{M}_0)^2}{\hat{M}_0} \leq \tilde{O}\left(\frac{1}{n}\right).$$

These inequalities further imply that with probability at least $1 - O(n^{-8})$,

$$\begin{aligned}
|T-1| &\le |\hat{M}_0 - M_0| + \sum_{\mu\ge 1} |\hat{M}_\mu - M_\mu| \\
&\le \tilde{O}\left(\frac{\sqrt{\Phi_1}+1}{n}\right) + \sum_{\mu\ge 1} \tilde{O}\left(\frac{\sqrt{\Phi_{\mu\mu}}}{n}\right) \\
&= \sum_{\mu\ge 0} \tilde{O}\left(\frac{\sqrt{\Phi_{\mu\mu}}}{n}\right) \\
&\le \tilde{O}\left(\sqrt{\frac{\sum_{\mu\ge 1} \mathbb{1}_{\Phi_\mu>0}}{n}}\right),
\end{aligned}$$

where the second inequality follows from $\sum_{\mu\ge 1} \mu\Phi_\mu < n\log n$ and the Cauchy-Schwarz inequality.

To characterize the performance of estimator $\hat{M}^* := \{\hat{M}^*_\mu\}_{\mu\ge 0}$, we bound the KL divergence by the χ-squared distance. By the above inequalities, with probability at least $1 - O(n^{-8})$,

$$\begin{aligned}
\sum_{\mu\ge 0} M_u \log\frac{M_\mu}{\hat{M}^*_\mu} &\le \sum_{\mu\ge 0} \frac{(M_\mu - \hat{M}^*_\mu)^2}{\hat{M}^*_\mu} \\
&\le 2(T-1)^2 + \sum_{\mu\ge 0} 2T\frac{(M_\mu - \hat{M}_\mu)^2}{\hat{M}_\mu} \\
&\le \tilde{O}\left(\frac{\sum_{\mu\ge 1}\mathbb{1}_{\Phi_\mu>0}}{n}\right) + \tilde{O}\left(\frac{1}{n}\right) + \sum_{\mu\ge 1}\tilde{O}\left(\frac{\mathbb{1}_{\Phi_\mu>0}}{n}\right) \\
&= \tilde{O}\left(\frac{\sum_{\mu\ge 1}\mathbb{1}_{\Phi_\mu>0}}{n}\right) \\
&= \tilde{O}\left(\frac{D_\Phi}{n}\right).
\end{aligned}$$

Finally, for each $x \in [k]$, define our probability estimate by

$$\hat{p}^*_x(X^n) = \frac{\hat{M}^*_{N_x}}{\Phi_{N_x}}.$$

The following identity [OS15] completes the proof of Theorem 1.

$$\tilde{\ell}_{X^n}(p,\hat{p}^*) = \sum_{\mu\geq 0} M_\mu \log \frac{M_\mu}{\hat{M}^*_\mu}.$$

3.C Proof of Corollary 9

We begin with a lemma that partially characterizes the shape of a log-concave distribution.

Lemma 24. *[DKS16a] Let p be a log-concave distribution with mean μ_p and standard deviation σ_p. Let $\alpha,\beta \in [k]$ be integers satisfying $\alpha \leq \mu_p - \Omega(\sigma_p(1+\log(1/\varepsilon)))$ and $\beta \geq \mu_p + \Omega(\sigma_p(1+\log(1/\varepsilon)))$. Then,*

$$\sum_{x=1}^{\alpha} p_x + \sum_{x=\beta}^{k} p_x \leq 2\varepsilon.$$

In addition, for σ_p larger than an absolute constant, the maximum probability satisfies

$$\max_{x\in[k]} p_x \in [1/(8\sigma_p), 1/\sigma_p].$$

Setting $\varepsilon = 1/n^5$ in the above lemma, we obtain

$$\begin{aligned}
\Pr(D_\Phi > \Omega(\log(n^5)\sigma)) &\leq \Pr(D > \Omega(\log(n^5)\sigma)) \\
&\leq \Pr(\exists x, s.t.\ x \notin (\alpha,\beta), N_x \geq 1) \\
&\leq \sum_{x=1}^{\alpha} np_x + \sum_{x=\beta}^{k} np_x \\
&\leq 2\cdot n^{-4}.
\end{aligned}$$

Therefore, $\mathbb{E}[D_\Phi] \leq O\left(\log(n^5)\sigma\right)$. Now, we use the second part of Lemma 9 to derive a different

upper bound on $\mathbb{E}[D_\Phi]$. Let $j_{\max}$ be the index such that $\max_{x\in[k]} p_x \in I_{j_{\max}}$.

$$(j_{\max}-1)^2\frac{\log n}{n} < \max_{x\in[k]} p_x < \frac{\log n}{\sigma}.$$

The above inequality implies $j_{\max} < \sqrt{2n/\sigma}+1$. Using the same reasoning as in Section 4.2 in the main paper, we get

$$\mathbb{E}[D_\Phi] \le O\left((\sqrt{n/\sigma})^{\frac{2}{3}}\cdot n^{\frac{1}{3}}\right)\cdot\log n = \tilde{O}\left((\sigma n)^{-\frac{1}{3}}\right).$$

Combining the above two upper bounds on $\mathbb{E}[D_\Phi]$ yields

Corollary 14. *For any distribution $p \in \mathcal{L}_k^{n,\sigma}$ and $p' \in \langle p\rangle$,*

$$\tilde{r}_n(p',\hat{p}^*) \le \tilde{O}\left((\sigma n)^{-\frac{1}{3}} \wedge \frac{\sigma}{n}\right).$$

3.D Proof of Corollary 11

Consider the collection $\mathcal{P}_k^{\alpha,c} := \{p\in\Delta_k : p_x \le c\cdot x^{-\alpha}\}$ of enveloped (truncated) power-law distributions. Note that this definition generalizes power-law families, and that distributions in $\mathcal{P}_k^{\alpha,c}$ are not necessarily log-convex. Let $\beta\in(0,1)$ be a parameter to be determined, and x_0 be the threshold such that $2n(c\cdot x_0^{-\alpha}) = n^\beta$. The symbols $x\in[k]$ that are no larger than x_0 contribute at most x_0 to D_Φ. On the other hand, for any $x > x_0$, we have $\mathbb{E}[N_x] = np_x \le n(c\cdot x^{-\alpha}) < 0.5n^\beta$. Therefore, for $x > x_0$,

$$\begin{aligned}
\Pr(N_x > 2n^\beta) &\le \frac{np_x}{1-p_x}\Pr(N_x \ge 2n^\beta)\\
&\le 2np_x\Pr\left(N_x \ge \mathbb{E}[N_x] + n^\beta\right)\\
&\le 2np_x\exp\left(-n^\beta/3\right),
\end{aligned}$$

where the first inequality follows from direct comparison and the last follows from the Chernoff bound for binomial random variables. By the union bound,

$$\begin{aligned}\Pr(\exists x \in [k]\ s.t.\ N_x > 2n^{\beta}) &\leq \sum_{x\in[k]} \Pr(N_x > 2n^{\beta}) \\ &\leq 2n\exp\left(-n^{\beta}/3\right).\end{aligned}$$

Therefore, with probability at least $1-2n\exp\left(-n^{\beta}/3\right)$,

$$D_{\Phi} \leq x_0 + 2n^{\beta} = (2c)^{\frac{1}{\alpha}} n^{\frac{1-\beta}{\alpha}} + 2n^{\beta}.$$

Optimizing the right-hand side by choosing $\beta = 1/(\alpha+1)$, the inequality simplifies to

$$D_{\Phi} \leq ((2c)^{\frac{1}{\alpha}} + 2) n^{\frac{1}{\alpha+1}}.$$

Since $D_{\Phi} \leq n$, we can convert this high-probability result into the expectation bound,

$$\mathbb{E}[D_{\Phi}] \leq O(n^{\frac{1}{\alpha+1}}).$$

Along with Corollary 6 in the main paper this implies

Corollary 15. *For any distribution $p \in \mathcal{P}_k^{\alpha,c}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \leq \tilde{O}_{c,\alpha}\left(n^{-\max\{\frac{\alpha}{\alpha+1}, \frac{1}{2}\}}\right).$$

Chapter 4

Data Amplification: Instance-Optimal Property Estimation

4.1 Introduction

Recent years have seen significant interest in estimating properties of distributions over large domains [VV11a, JVHW15, JHW16, WY16, OSW16, ADOS17, HOSW18, WY19, HO19a, HO19c, CSS19b, HL20a]. Chief among these properties are support size and coverage, Shannon entropy, and L_1 distance to a known distribution. The main achievement of these papers is essentially estimating properties of distributions with alphabet size k using just $k/\log k$ samples.

In practice however, the underlying distributions are often simple, and their properties can be accurately estimated with significantly fewer than $k/\log k$ samples. For example, if the distribution is concentrated on a small part of the domain, or is exponential, very few samples may suffice to estimate the property. To address this discrepancy, [HOSW18] took the following competitive approach.

The best-known distribution property estimator is the *empirical estimator* that replaces the unknown underlying distribution by the observed empirical distribution. For example, with

n samples, it estimates entropy by the formula $-\sum_i (N_i/n)\log(N_i/n)$ where N_i is the number of times symbol i appeared. Besides its simple and intuitive form, the empirical estimator is also consistent, stable, and universal. It is therefore the most commonly used property estimator for data-science applications.

The estimator in [HOSW18] uses n samples and for any underlying distribution achieves the same performance that the empirical estimator would achieve with $n\sqrt{\log n}$ samples. It therefore provides an effective way to *amplify* the amount of data available by a factor of $\sqrt{\log n}$, regardless of the domain or structure of the underlying distribution.

In this paper we present novel estimators that increase the amplification factor for all sufficiently smooth properties including those mentioned above from $\sqrt{\log n}$ to the information-theoretic bound of $\log n$. Namely, for *every* distribution their expected estimation error with n samples is that of the empirical estimator with $n\log n$ samples and no further uniform amplification is possible.

It can further be shown [VV11a, JVHW15, ADOS17, WY19] that the empirical estimator estimates all of the aforementioned four properties with linearly many samples, hence the sample size required by the new estimators is always at most the $k/\log k$ guaranteed by the state-of-the-art estimators.

The current formulation has several additional advantages over previous approaches, which we illustrate as follows.

Fewer assumptions It eliminates the need for some commonly used assumptions. For example, support size cannot be estimated with any number of samples, as arbitrarily-many low-probabilities may be missed. Hence previous research [ADOS17, WY19] unrealistically assumed prior knowledge of the alphabet size k, and additionally that all positive probabilities exceed $1/k$. By contrast, the current formulation does not need these assumptions. Intuitively, if a symbol's probability is so small that it won't be detected even with $n\log n$ samples, we do not need to worry about it.

Refined bounds For some properties, our results are more refined than previously shown. For example, existing results estimate the support size to within $\pm\varepsilon k$, rendering the estimates rather inaccurate when the true support size S is much smaller than k. By contrast, the new estimation errors are bounded by $\pm\varepsilon S$, and are therefore accurate regardless of the support size. A similar improvement holds for the support coverage that we introduce below.

Graceful degradation For the previous results to work, one needs at least $k/\log k$ samples. With fewer samples, the estimators have no guarantees. By contrast, the guarantees of the new estimators work for any sample size n. If $n < k/\log k$, the performance may degrade, but will still track that of the empirical estimators with a factor $\log n$ more samples. See Theorem 1 for an example.

Instance optimality With the recent exception of [HOSW18], all modern property-estimation research took a min-max-related approach, evaluating the estimation improvement based on the worst possible distribution for the property. In reality, practical distributions are rarely the worst possible and often quite simple, rendering min-max approach overly pessimistic, and its estimators, typically suboptimal in practice. In fact, for this very reason, practical distribution estimators do not use min-max based approaches [GS95]. By contrast, our *competitive*, or *instance-optimal*, approach provably ensures amplification for every underlying distribution, regardless of its complexity or support size.

In addition, the proposed estimators run in time near-linear in the sample size, and the constants involved are very small, attributes shared by some, though not all existing estimators.

Below, we formalize the foregoing discussion in definitions.

Let Δ_k denote the collection of discrete distributions over $[k] := \{1,\ldots,k\}$. A distribution *property* is a mapping $F : \Delta_k \to \mathbb{R}$. It is *additive* if it can be written as

$$F(p) := \sum_{i\in[k]} f_i(p_i),$$

where $f_i : [0,1] \to \mathbb{R}$ are real functions. Many important distribution properties are additive:

Shannon entropy $H(p) := \sum_{i\in[k]} -p_i \log p_i$, is the principal measure of information and randomness [CT12], and arises in many machine-learning [CL68, QKC13, Bre15], neuroscience [MS95, VSLS+97, GK02], and other applications.

L_1 distance $D_q(p) := \sum_{i\in[k]} |p_i - q_i|$, where q is a given distribution, is one of the most basic and well-studied properties in the field of distribution property testing [BFR+00, Ron10, VV16, Can17].

Support size $S(p) := \sum_{i\in[k]} \mathbb{1}_{p_i>0}$, is a fundamental quantity for discrete distributions, and plays an important role in vocabulary size [McN73, ET76, TE87] and population estimation [Goo53, ML07].

Support coverage $C(p) := \sum_{i\in[k]}(1-(1-p_i)^m)$, for a given m, represents the number of distinct elements we would expect to see in m independent samples, arises in many ecological [Cha84, CL92, CCG+12, CC14], biological [Cha84, KLR99], genomic [ILLL09] as well as database [HNSS95] studies.

4.2 Prior and New Results

Given an additive property F and sample access to an unknown distribution p, we would like to estimate the value of $F(p)$ as accurately as possible. Let $[k]^n$ denote the collection of all length-n sequences, an estimator is a function $\hat{F} : [k]^n \to \mathbb{R}$ that maps a sample sequence $X^n \sim p$ to a property estimate $\hat{F}(X^n)$. We evaluate the performance of $\hat{F}$ in estimating $F(p)$ via its *mean absolute error* (MAE) [1],

$$L_{\hat{F}}(p,n) := \mathop{\mathbb{E}}_{X^n\sim p} \left|\hat{F}(X^n) - F(p)\right|.$$

[1] As we aim to estimate only a single property value, the estimators in this paper all have negligible variances, e.g., $O(1/n^{0.9})$. Hence the MAE is the same as MSE for our purpose, and we choose the former because it induces cleaner expressions.

Since we do not know p, the common approach is to consider the worst-case MAE of $\hat{F}$ over Δ_k,

$$L_{\hat{F}}(n) := \max_{p \in \Delta_k} L_{\hat{F}}(p, n).$$

The best-known and most commonly-used property estimator is the *empirical plug-in estimator*. Upon observing X^n, let N_i denote the number of times symbol $i \in [k]$ appears in X^n. The empirical estimator estimates $F(p)$ by

$$\hat{F}^{\mathrm{E}}(X^n) := \sum_{i \in [k]} f_i\left(\frac{N_i}{n}\right).$$

Starting with Shannon entropy, it has been shown [WY16] that for $n \geq k$, the worst-case (max) MAE of the empirical estimator $\hat{H}^{\mathrm{E}}$ is

$$L_{\hat{H}^{\mathrm{E}}}(n) = \Theta\left(\frac{k}{n} + \frac{\log k}{\sqrt{n}}\right). \tag{4.1}$$

On the other hand, [JVHW15, WY16, ADOS17, HO19a, HO19c] showed that for $n \geq k/\log k$, more sophisticated estimators achieve the best min-max performance of

$$L(n) := \min_{\hat{F}} L_{\hat{F}}(n) = \Theta\left(\frac{k}{n \log n} + \frac{\log k}{\sqrt{n}}\right). \tag{4.2}$$

Hence up to constant factors, for the "worst" distributions, the MAE of these estimators with n samples equals that of the empirical estimator with $n \log n$ samples. A similar relation holds for the other three properties we consider.

However, the min-max formulation is pessimistic as it evaluates the estimator's performance for the worst distributions. In many practical applications, the underlying distribution is fairly simple and does not attain this worst-case loss, rather, a much smaller MAE can be achieved. Several recent works have therefore gone beyond worst-case analysis and designed algorithms that perform well for all distributions, not just those with the worst performance [OS15, VV16, HO19b].

For property estimation, [HOSW18] designed an estimator $\hat{F}^{\mathrm{A}}$ that for any underlying distribution uses n samples to achieve the performance of the $n\sqrt{\log n}$-sample empirical estimator, hence effectively multiplying the data size by a $\sqrt{\log n}$ *amplification factor*.

Lemma 1 ([HOSW18])**.** *For every property F in a large class including the aforementioned four properties, there is an absolute constant c_F such that for all distributions p, all $\varepsilon \le 1$, and all $n \ge 1$,*

$$L_{\hat{F}^{\mathrm{A}}}(p,n) - L_{\hat{F}^{\mathrm{E}}}\left(p, \varepsilon n\sqrt{\log n}\right) \le c_F \cdot \varepsilon.$$

In this work, we fully strengthen the above result and establish the limits of data amplification for all sufficiently smooth additive properties including four of the most important ones, and all that are appropriately Lipschitz.

Using Shannon entropy as an example, we achieve a $\log n$ amplification factor. Equations (4.1) and (4.2) imply that the improvement over the empirical estimator cannot always exceed $O(\log n)$, hence up to an absolute constant, this amplification factor is information-theoretically optimal. Similar optimality arguments hold for our results on the other three properties (e.g., see Table 1 in [ADOS17]).

Specifically, we derive efficient estimators $\hat{H}$, $\hat{D}$, $\hat{S}$, $\hat{C}$, and $\hat{F}$ for the Shannon entropy, L_1 distance, support size, support coverage, and a broad class of additive properties which we refer to as *Lipschitz properties*. These estimators run in *near-linear time*, take a single parameter ε, and given samples $X^n \sim p$, amplify the data as described below.

For brevity, henceforth we shall write $a \wedge b$ and $a \lesssim b$ instead of $\min\{a,b\}$ and $a = O(b)$, respectively, and abbreviate support size $S(p)$ by S_p and coverage $C(p)$ by C_p.

The following five theorems hold for all $\varepsilon \le 1$, all distributions p, and all $n \ge 1$.

Theorem 1 (Shannon entropy)**.**

$$L_{\hat{H}}(p,n) - L_{\hat{H}^{\mathrm{E}}}(p, \varepsilon n \log n) \lesssim \varepsilon \wedge \left(\frac{S_p}{n} + \frac{1}{n^{0.49}}\right).$$

Note that the estimator requires no knowledge of S_p or k. When $\varepsilon = 1$, the estimator amplifies the data by a factor of $\log n$. As ε decreases, the amplification factor decreases, and so does the extra additive inaccuracy. One can also set ε to be a vanishing function of n, e.g., $\varepsilon = 1/\log\log n$.

This result may be interpreted as follows. For distributions with large support sizes such that the min-max estimators provide no or only very weak guarantees, our estimator with n samples always tracks the performance of the $n\log n$-sample empirical estimator. On the other hand, for distributions with relatively small support sizes, our estimator achieves a near-optimal $O(S_p/n)$-error rate.

Similarly, for L_1 distance to a fixed distribution q,

Theorem 2 (L_1 distance)**.** *For any q, we can construct an estimator $\hat{D}_q$ for D_q such that*

$$L_{\hat{D}_q}(p,n) - L_{\hat{D}_q^{\mathrm{E}}}\left(p, \varepsilon^2 n\log n\right) \lesssim \varepsilon \wedge \left(\sqrt{\frac{S_p}{n}} + \frac{1}{n^{0.49}}\right).$$

Besides having an interpretation similar to that of Theorem 1, the above result shows that for each q and each p, we can use just n samples to achieve the performance of the $n\log n$-sample empirical estimator. More generally, for any additive property $F(p) := \sum_{i\in[k]} f_i(p_i)$ that satisfies the simple condition: f_i is $O(1)$-Lipschitz, for all i,

Theorem 3 (General additive properties)**.** *Given F, we can construct an estimator $\hat{F}$ such that*

$$L_{\hat{F}}(p,n) - L_{\hat{F}^{\mathrm{E}}}\left(p, \varepsilon^2 n\log n\right) \lesssim \varepsilon \wedge \left(\sqrt{\frac{S_p}{n}} + \frac{1}{n^{0.49}}\right).$$

The results in [KOPS15] show that no plug-in estimators provide those theoretical guarantees presented in Theorem 2 and 3. Henceforth, we refer to the above collection of distribution properties as the class of *Lipschitz properties*. The L_1 distance D_q, for any q, is in this class.

Lipschitz properties are essentially bounded by absolute constants and Shannon entropy

grows at most logarithmically in the support size, and we were able to approximate all with just an additive error. Support size and support coverage can grow linearly in k and m, and can be approximated only multiplicatively. We therefore evaluate the estimator's normalized performance, regarding the property value.

Note that for both properties, the amplification factor is logarithmic in the property value, which can be arbitrarily larger than the sample size n.

The following two theorems hold for $\varepsilon \leq e^{-2}$,

Theorem 4 (Support size)**.**

$$\frac{1}{S_p}\left(L_{\hat{S}}(p,n) - L_{\hat{S}^{\mathrm{E}}}\left(p, n\cdot\frac{\log S_p}{\log^2\varepsilon}\right)\right) \lesssim \varepsilon + S_p^{\frac{1}{|\log\varepsilon|}-\frac{1}{2}}.$$

To make the slack term vanish, one can simply set ε to be a vanishing function of n (or S_p), e.g., $\varepsilon = 1/\log n$. Note that in this case, the slack term modifies the multiplicative error in estimating S_p by only $o(1)$, which is negligible in most applications. Similarly, for support coverage,

Theorem 5 (Support coverage)**.**

$$\frac{1}{C_p}\left(L_{\hat{C}}(p,n) - L_{\hat{C}^{\mathrm{E}}}\left(p, n\cdot\frac{\log C_p}{\log^2\varepsilon}\right)\right) \lesssim \varepsilon + C_p^{\frac{1}{|\log\varepsilon|}-\frac{1}{2}}.$$

The next section presents implications of these results.

4.3 Implications

Data amplification Numerous modern scientific applications, such as those emerging in social networks and genomics, deal with properties of distributions whose support size S_p is equal to or even larger than the sample size n.

In this data-sparse regime, the estimation error of the empirical estimator often decays at a slow rate, e.g., $1/\log^c n$ for some $c \in (0,1)$, hence the proposed estimators yield a much more

accurate estimate, paralleling that of the empirical with $n \log n$ samples. For applications where $n \geq 25,000$ and regardless of the distribution structure, our approach significantly amplifies the number of samples by at least a factor of 10, known by practitioners as an "order of magnitude".

As for the data-rich regime where $n \gg S_p$, our method essentially recovers the the standard $\sqrt{S_p/n}$ rate of maximum likelihood methods in general, without knowing S_p.

Instance optimality With just n samples, our method emulates the performance of the $n \log n$-sample empirical estimator for *every distribution instance*. The method hence possesses the vital ability of strengthening all MAE guarantees of the empirical estimator by a logarithmic factor, which is optimal in many settings.

The significance of such "instance optimality" arises from 1) empirical estimators are often simple and easy to analyze; 2) there is a rich literature on their estimation attributes, e.g., [Bus17] and the references therein; 3) empirical estimators are the best-known and most-used.

Consequently, we can work on a simple problem, analyzing the performance of the empirical estimator, and immediately strengthen the result we obtain by a logarithmic-factor using the theorems in this paper. In many cases, the strengthened results are challenging to establish via other statistical methods. We present two examples below.

Entropy Consider entropy estimation over Δ_k. As Equation 4.2 shows, the min-max MAE is known for $n \geq k/\log k$, and essentially becomes a constant when n gets close to the $k/\log k$ lower bound. Nevertheless, over an alphabet of size k, the value of entropy can go up to $\log k$. Hence, it is still possible to get meaningful estimation results in the $n = o(k/\log k)$ large-alphabet regime.

We follow the above strategy to solidify the statement. First, for empirical estimator $\hat{H}^{\mathrm{E}}$, [Pan03] [see Proposition 1] provides a short argument showing that its worst-case MAE, for all n and k, satisfies

$$L_{\hat{H}^{\mathrm{E}}}(n) \leq \log\left(1 + \frac{k-1}{n}\right) + \frac{\log n}{\sqrt{n}}.$$

Consolidating this inequality with Theorem 1 then implies

Corollary 1. *In the $n = o(k/\log k)$ large-alphabet regime, the min-max MAE of estimating Shannon entropy satisfies*

$$L(n) \le (1+o(1))\log\left(1+\frac{k-1}{n\log n}\right).$$

Lipschitz Property The same type of arguments apply to any Lipschitz property F. Again, we begin with characterizing the performance of the empirical estimator $\hat{F}^{\mathrm{E}}$. By Lemma 3 and the Cauchy-Schwarz inequality, the bias of $\hat{F}^{\mathrm{E}}$ is at most $O(\sqrt{k/n})$. By the Efron-Stein inequality, the standard deviation is no more than $O(1/\sqrt{n})$.

It then follows by Theorem 3 that: $\hat{F}$ estimates F over Δ_k to an MAE of ε with $O(k/(\varepsilon^3\log k))$ samples. Note that 1) this yields the first estimator for Lipschitz properties with optimal sample dependence on k; 2) after a draft of this paper became available online, [HO19c] improved the sample dependence on ε to the optimal ε^2.

4.4 Estimator Construction and Analysis

For clarity, we focus on the proof of Theorem 1 about entropy estimation, and explain only necessary modifications for similar arguments to go through for other properties. We begin by relating the empirical entropy estimator to the "Bernstein polynomial" of function $-x\log x$.

Notation For a sampling parameter n and accuracy $\varepsilon \le 1$, define the *amplification factor* as $a := \varepsilon\log n$. Without loss of generality, assume that $\varepsilon \ge 1/\log n$ and hence $a \ge 1$. For simplicity, write $h(x) := -x\log x$, $m := na$, $\tau_n := \Theta(\log n/n)$ and $d_n := \Theta(\log n)$, where the asymptotic notations hide only properly chosen absolute constants.

4.4.1 Bernstein Polynomial

Drawing i.i.d.samples Y^m from any distribution p, the expected value of the empirical estimator for entropy is

$$\mathbb{E}[\hat{H}^{\mathrm{E}}(Y^m)] = \sum_{i\in[k]} \mathop{\mathbb{E}}_{M_i\sim\mathrm{bin}(m,p_i)} \left[h\left(\frac{M_i}{m}\right)\right].$$

Note that for any function f, $m\in\mathbb{N}$, and $x\in[0,1]$, the degree-m *Bernstein polynomial* of f is

$$B_m(f,x) := \sum_{j=0}^{m} f\left(\frac{j}{m}\right)\binom{m}{j} x^j(1-x)^{m-j}.$$

Therefore, we can express the expectation of the empirical entropy estimator as

$$\mathop{\mathbb{E}}_{Y^m\sim p}[\hat{H}^{\mathrm{E}}(Y^m)] = \sum_{i\in[k]} B_m(h,p_i).$$

As modifying a sample changes the value of $\hat{H}^{\mathrm{E}}(Y^m)$ by at most $2\log m/m$, the Efron-Stein inequality bounds its variance by $2\log^2 m/m$, which is negligible in magnitude. Hence, for our purpose, we focus on finding a good approximation of each $B_m(h,p_i)$.

4.4.2 Estimator Construction and Computation

In the subsequent sections, given i.i.d.samples $X^n\sim p$, we construct our estimator as follows.

Substitute n by $2n$ for simplicity. According to Section 4.4.4, we first split the samples into two halves, X_1^n and X_{n+1}^{2n}, and respectively denote by N_i and N_i' the empirical counts of each symbol $i\in[k]$ in them.

Then, we follow [Dob58] to classify the symbols into two categories and decompose the

sum

$$\mathop{\mathbb{E}}_{Y^m\sim p}[\hat{H}^{\mathrm{E}}(Y^m)] = \sum_{i\in[k]} B_m(h,p_i)$$

into two parts by thresholding the empirical counts N_i' at level $1/\varepsilon$. The first part, $\mathcal{V}_L := \sum_{i\in[k]} B_m(h,p_i)\mathbb{1}_{N_i'>1/\varepsilon}$, corresponds to the contribution of symbols with potentially large probabilities. Illustrated in Section 4.4.3, this quantity is well approximated by the *large-probability estimator*

$$\hat{\mathcal{V}}_L := \sum_{i\in[k]} h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N_i'>\frac{1}{\varepsilon}},$$

to an MAE of $2(\varepsilon\wedge S_p/n)$. As for the small-probability part,

$$\mathcal{V}_S := \sum_{i\in[k]} B_m(h,p_i)\cdot \mathbb{1}_{N_i'\le\frac{1}{\varepsilon}},$$

we follow the arguments in Section 4.4.4 and 4.4.5 to learn each summand adaptively (to the magnitude of the probability) and compute the summation.

Concretely, recall $\tau_n = \Theta(\log n/n)$ and $d_n = \Theta(\log n)$. For a given function and domain, the polynomial achieving the minimal maximum deviation from the function over the domain is the *min-max polynomial*. Then, denote by

$$\tilde{h}_m(x) := \sum_{t=0}^{d_n} b_t x^t$$

the degree-d_n min-max polynomial of $B_m'(h,p_i)$ over interval $I_n := [0,\tau_n]$. The *small-probability estimator* is

$$\hat{\mathcal{V}}_S := \sum_{i\in[k]}\left(\sum_{t=1}^{d+1}\frac{b_{t-1}}{t}\cdot\frac{N_i^t}{n^t}\right)\cdot\mathbb{1}_{N_i\lesssim\log n}\cdot\mathbb{1}_{N_i'\le\frac{1}{\varepsilon}},$$

where for each symbol i, the term in the parentheses is an unbiased estimator for $\tilde{H}_m(p_i) := \int_0^{p_i}\tilde{h}_m(s)ds$. Next, we illustrate the technique and intuition behind the construction.

Differential smoothing The construction of $\hat{\mathcal{V}}_S$ presents a generic method for designing

a polynomial $\tilde{G}$ that closely approximates a given differentiable function G *with pointwise error bounds*.

More precisely, for a fixed interval $I := [0, \tau]$ and degree bound $d \in \mathbb{N}$, we want to find a polynomial $\tilde{G}$ of degree at most d, satisfying

$$\max_{x \in I} |\tilde{G}(x) - G(x)| \leq c \cdot x,$$

for a number $c \geq 0$ that is *as small as possible*.

We propose a novel method, *differential smoothing*, that addresses this approximation problem and operates as follows.

1. Compute $G'(x)$ and write $g := G'$.
2. Approximate g by its min-max polynomial $\tilde{g}$ over I.
3. Let c be the min-max approximation error in Step 2.
4. Compute $\tilde{G}(x) := \int_0^x \tilde{g}(t)dt$.

By the triangle inequality for integrals, the resulting c and $\tilde{G}$ satisfy the desired inequality. Besides, Step 2 and 3 can be jointly performed using the well-known Remez algorithm [PT09, Tre13].

Turning back to our estimator $\hat{V}_S$, by the reasoning in Section 4.4.6 and 4.4.7, the min-max polynomial $\tilde{h}_m(x)$ approximates $B'_m(h,x)$ to within $O(\varepsilon)$ over I_n. Hence, applying the method of differential smoothing yields

$$|B_m(h,x) - \tilde{H}_m(x)| \lesssim \varepsilon \cdot x.$$

Further relating this inequality to the expectation of the empirical entropy estimator implies

$$\left| \mathop{\mathbb{E}}_{Y^m \sim p} [\hat{H}^{\mathrm{E}}(Y^m)] - \sum_{i \in [k]} \tilde{H}_m(p_i) \right| \lesssim \sum_{i \in [k]} \varepsilon \cdot p_i = \varepsilon.$$

In Section 6.1 of the supplementary, we prove that the absolute bias is also at most $O(S_p/n)$, which requires some additional work. Finally, Section 7.1 bounds the mean absolute deviation of the estimator by $O(1/n^{0.49})$.

Consequently, we approximate $H(p)$ by

$$\hat{H} := \hat{\mathcal{V}}_L + \hat{\mathcal{V}}_S.$$

Computational complexity The dominant computation step is finding the min-max polynomial of $B'_m(h,x)$, for which we utilize the well-known Remez algorithm [PT09, Tre13]. In Section 9 of the supplementary, we shall argue that the algorithm takes only $\tilde{O}(n)$ time to well approximate $B'_m(h,x)$.

4.4.3 Large-Probability Estimator

Following the previous arguments, we say that $i \in [k]$ is a *large-probability symbol* if $N'_i > 1/\varepsilon$. To the expectation of the m-sample empirical estimator, these symbols contribute

$$\mathcal{V}_L = \sum_{i\in[k]} B_m(h,p_i)\cdot \mathbb{1}_{N'_i > \frac{1}{\varepsilon}}.$$

We estimate $\mathcal{V}_L$ by respectively reweighing the empirical estimator associated with the first-half samples:

$$\hat{\mathcal{V}}_L = \sum_{i\in[k]} h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N'_i > \frac{1}{\varepsilon}}.$$

To bound the estimation bias, we leverage the next lemma, stating that the Bernstein polynomial of h closely approximates the function over $[0,1]$.

Lemma 2. *For any $t \in \mathbb{Z}^+$ and $x \in [0,1]$,*

$$-\frac{1-x}{t} \le B_t(h,x) - h(x) \le 0.$$

The number of symbols satisfying $N_i' > 1/\varepsilon$ is at most $n\varepsilon$. Together with the lemma and triangle inequality, this yields

$$|\mathbb{E}[\mathcal{V}_L] - \mathbb{E}[\hat{\mathcal{V}}_L]| \leq \sum_{i\in[k]} \left(\frac{1+a}{m}\right)(1-p_i)\mathbb{E}\left[\mathbb{1}_{N_i'>\frac{1}{\varepsilon}}\right] \leq 2\varepsilon.$$

Furthermore, the number of such symbols is also at most S_p, implying an alternative upper bound of $2S_p/n$.

For Shannon entropy, we note that adding $1/(2n)$ to the empirical estimate $h(N_i/n)$ may reduce its bias. This particular method, known as the "Miller-Mallow estimator", appears in [Mil55] and eliminates the first-order term of $B_n(h,x) - h(x)$. Applying the method will introduce extra complications in the analysis, and hence for entropy and general non-differentiable properties, we employ the original empirical estimator. On the other hand, substituting the Miller-Mallow estimate into our algorithm in Theorem 1 retains its theoretical guarantee.

For Lipschitz properties, the rich literature on Bernstein operators presents us with the following bound.

Lemma 3 ([Bus17] Proposition 4.9)**.** *For any $t \in \mathbb{Z}^+$, $x \in [0,1]$, and c-Lipschitz function f,*

$$|B_t(f,x) - f(x)| \leq c \cdot \sqrt{\frac{x(1-x)}{t}}.$$

Combined with the Cauchy-Schwarz inequality, the lemma shows that the estimation bias of the respective $\hat{\mathcal{V}}_L$ admits

$$|\mathbb{E}[\mathcal{V}_L] - \mathbb{E}[\hat{\mathcal{V}}_L]| \leq 2\left(\varepsilon \wedge \sqrt{\frac{S_p}{n}}\right).$$

This completes the bias analysis of the large-probability estimator, while Section 6.2 in the supplementary provides additional technical details. For the variance analysis, see Section 7.2. The following three sections proceed to construct the small-probability estimator and introduce

fundamental results from polynomial approximation theory.

4.4.4 Choice of Parameters and Sample Splitting

Section 4.4.1 calls for estimating $B_m(h,x)$. Applying the method of *differential smoothing* in Section 4.4.2, we first choose some domain $I = [0,\tau]$ and degree d, and estimate $B'_m(h,x)$ by its min-max polynomial $\tilde{h}_m(x) = \sum_{t=0}^{d} b_t x^t$ over I. Then, we approximate $B_m(h,x)$ by

$$\tilde{H}_m(x) = \int_0^x \tilde{h}_m(t)dt = \sum_{t=0}^{d} \frac{b_t}{t+1} x^{t+1}.$$

To estimate $\tilde{H}_m(x)$, note that given a binomial variable $X \sim \text{bin}(n,x)$, an unbiased estimator for x^t is $X^{\underline{t}}/n^{\underline{t}}$, where $t \in \mathbb{N}$ and $A^{\underline{B}}$ denotes *the B-th order falling factorial* of A. Hence, we employ

$$\hat{H}_m(X) := \sum_{t=1}^{d+1} \frac{b_{t-1}}{t} \cdot \frac{X^{\underline{t}}}{n^{\underline{t}}},$$

an *unbiased* estimator for $\tilde{H}_m(x)$ that corresponds to the parenthetical component in estimator $\hat{\mathcal{V}}_S$'s expression. Next, we illustrate the intuitions behind our choice of τ and d.

For any $X \sim \text{bin}(n,x)$, the variance of $\hat{H}_m(X)$ generally gets larger as the degree d increases. On the other hand, a higher-degree polynomial is able to achieve a lower approximation error. To balance this bias-variance trade-off, we want to reduce both the interval length, τ, and the polynomial degree, d, while maintaining the approximation power.

As in Section 4.4.2, we set parameter $\tau = \tau_n = \Theta(\log n/n)$ since below this threshold, sample statistics are insufficient for inferring the relative magnitudes of the underlying probabilities with high confidence. Regarding the degree parameter $\tau = \tau_n = \Theta(\log n)$, below the $\log n$ threshold, the approximation $\tilde{H}_m$ loses the $\varepsilon \cdot x$ guarantee; in contrast, above the threshold, the final estimator may no longer possess a vanishing variance. For more details, see derivations in Section 7.1 and Appendix A of the supplementary.

One thing that follows the construction of $\tilde{H}_m$ and $\hat{H}_m$ is how to apply these approximations to only probabilities of order τ_n. This issue arises from the fact that we observe symbol counts, not ranges of the actual probability values. It is straightforward to deal with such uncertainty by inferring the magnitudes of unknowns leveraging the counting statistics concentration.

For concentration, binomial random variables are sums of independent indicator variables and possess Gaussian-type tail bounds. To avoid introducing additional statistical dependency, we 1) split the sample sequence into two halves of equal length; 2) denote respectively the empirical counts of each symbol i in the first and second halves by N_i and N_i' (where we slightly abused the notation); 3) classify each $i \in [k]$ as a large- or small- probability symbol by thresholding the count N_i' at $1/\varepsilon$. The supplementary material presents relevant details in Section 5 and 6.2.

In the literature, the above procedure is often referred to as *sample splitting*. This idea of classifying the symbols in the alphabet into two categories dates back to [Dob58], and has been applied to estimate a variety of specific distribution properties in the past decade [AOST14, JVHW15, WY16, HOSW18]. Recently, [HO19c] generalize this idea to estimate general properties by partitioning the unit interval into $\tilde{\Theta}(\sqrt{n})$ pieces; [HO19b] apply the method to derive state-of-the-art distribution estimators.

Sample splitting and additiveness of the property enable us to estimate the contributions from the large and small probabilities separately. The rest sections assume this separation and address the small-probability approximation error.

4.4.5 Min-Max Polynomial

Polynomials have extensive applications to statistical inference, ranging from approximating the norms of Gaussian parameters [CL11] to learning structured distributions [CDSS14, ADLS17, HO19b] to estimating properties of distributions [JVHW15, OSW16, WY16, HOSW18, HO19c].

As illustrated in Section 4.4.2 and 4.4.4, we aim to find a polynomial $\tilde{h}_m(x)$ of degree

$d_n = \Theta(\log n)$ that satisfies the pointwise bound $|B'_m(h,x) - \tilde{h}_m(x)| \lesssim \varepsilon$ over $I_n = [0, \tau_n]$.

The task naturally calls for a polynomial achieving the minimal maximum deviation from $B'_m(h,x)$, commonly known as the respective *min-max polynomial*. Moreover, direct computation shows that $B'_m(h,x)$ is the order-$(m-1)$ Bernstein polynomial of another function:

$$B'_m(h,x) = B_{m-1}(h_m, x),$$

where function h_m is defined as

$$h_m(y) := -\log\frac{m-1}{m} + (m-1)\left(h\left(y + \frac{1}{m-1}\right) - h(y)\right).$$

Hence, our objective reduces to bounding the error of min-max polynomial approximations of $B_{m-1}(h_m, x)$ over I_n. As one could expect, the analysis gets more involved since 1) $B_{m-1}(h_m, x)$ is a high-degree polynomial with transcendental coefficients; 2) in general, there are no closed-form formulas for the min-max polynomials of a real function.

Though sophisticated in its form, function $B_{m-1}(h_m, x)$ is continuous and relatively smooth, as hinted by Lemma 2. This simple observation serves as the starting point for our subsequent analysis. In the next section, we dive into approximation theory and present fundamental connections between the smoothness of a function (characterized by specific quantities) and its min-max polynomial approximation error over a given interval. The desired result then follows by a sequence of inequalities and simplifications that enable us to gauge the smoothness of $B_{m-1}(h_m, x)$.

For the proof of the derivative identity on h_m and a more straightforward argument leading to a weaker result, see Section 4 and 5 of the supplementary.

4.4.6 Moduli of Smoothness

In this section, we introduce some notable results in approximation theory [DT12] that are crucial for simplifying the problem. Denote $\varphi(x) := \sqrt{x(1-x)}$. For any function $f : [0,1] \to \mathbb{R}$, *the first- and second- order Ditzian-Totik moduli of smoothness* quantities of f are

$$w_\varphi^1(f,t) := \sup\left\{|f(u)-f(v)| : 0 \le u,v \le 1, \right. \\ \left. |u-v| \le t \cdot \varphi\left(\frac{u+v}{2}\right)\right\},$$

and

$$w_\varphi^2(f,t) := \sup\left\{\left|f(u)+f(v)-2f\left(\frac{u+v}{2}\right)\right| : \right. \\ \left. 0 \le u,v \le 1, |u-v| \le 2t \cdot \varphi\left(\frac{u+v}{2}\right)\right\},$$

respectively. Let P_d denote the collection of polynomials with real coefficients and degree at most d. For any $d \in \mathbb{Z}^+$, interval $I \subset \mathbb{R}$, and function $f : I \to \mathbb{R}$, denote by

$$E_d[f,I] := \min_{\tilde{f} \in P_d} \max_{x \in I} |f(x) - \tilde{f}(x)|$$

the *best approximation error* of the degree-d min-max polynomial of f over I. For a bounded domain I, we can always shift and rescale f to make it a real function over $[0,1]$. Hence, without loss of generality, it suffices to consider and analyze $E_d[f] := E_d[f,[0,1]]$.

The connection between the best polynomial-approximation error $E_d[f]$ of a continuous function f and the second-order Ditzian-Totik moduli of smoothness $w_\varphi^2(f,t)$ is established in the following lemma [DT12].

Lemma 4. *There are absolute constants C_1 and C_2 such that for any continuous function f over*

$[0,1]$ *and* $d > 2$,

$$E_d[f] \leq C_1 w_\varphi^2(f, d^{-1}),$$

and

$$\frac{1}{d^2}\sum_{t=0}^{d}(t+1)E_t[f] \geq C_2 w_\varphi^2(f, d^{-1}).$$

The above lemma shows that the second-order smoothness quantity $w_\varphi^2(f,\cdot)$ essentially characterizes $E_\cdot[f]$, and thus transforms the problem of showing

$$|\tilde{h}_m(x) - B_{m-1}(h_m, x)| \lesssim \varepsilon, \forall x \in I_n,$$

to that of establishing

$$w_\varphi^2(B_{m-1}(h_m, \tau_n \cdot x), d_n^{-1}) \lesssim \varepsilon,$$

where $\tau_n = \Theta(\log n/n)$ and $d_n = \Theta(\log n)$ by definition.

4.4.7 Simplification via Poissonization

The last block in our analysis is Poissonization, which helps decompose and simplify the function to approximate. For any $y \in [0,\infty]$, define two functions:

$$f_1(y) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(y)}[h(X)] = -e^{-y}\sum_{j=1}^{\infty}\frac{y^j}{j!}(j\log j)$$

and

$$f_2(y) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(y)}[h(X+1)].$$

Let $z(x) := (m-1)x$ for simplicity. The following lemma, appearing in Appendix A.1 of the supplementary relates $B_{m-1}(h_m, x)$ to these functions and base function $h(x)$.

Lemma 5. *For any $m \in \mathbb{Z}^+$ and $x \in [0, \log^4 m/m]$,*

$$h_m(x) - B_{m-1}(h_m, x) = [h(z(x)+1) - f_2(z(x))] \\ - [h(z(x)) - f_1(z(x))] + \tilde{O}\left(\frac{1}{m}\right).$$

In particular, the above equation holds for any sufficiently large n and $x \in I_n = [0, \tau_n]$. Since $1/m = 1/(na-1) \leq \min\{1/\log n, S_p/n\}$, the last term on the right-hand side is negligible. These results, together with the function-wise triangle inequality on w_φ^2, further reduce the last inequality in Section 4.4.6 to bounds in the form of

$$w_\varphi^2(g(x), d_n^{-1}) \lesssim \varepsilon,$$

for function $g(x)$ being $h_m(\tau_n \cdot x)$, $h(z(x))$, $h(z(x)+1)$, $f_1(z(x))$, and $f_2(z(x))$, respectively.

We prove these bounds in Appendix A.2 and A.3 of the supplementary. In Appendix B, a similar yet more involved argument extends the result to all Lipschitz properties. One reason for the extra complication is the absence of concrete expression, as we impose only the Lipschitz condition.

While these proofs are technical, a critical insight is that the optimization problems induced by computing w_φ^2 for the above choices of g are all convex. Consequently, it suffices to consider only the boundary cases of parameters.

4.5 Experiments

We demonstrate the efficacy of the proposed estimators by comparing their performance to two state-of-the-art estimators [WY16, WY19], and empirical estimators with logarithmic larger sample sizes. Due to method similarity, we present only the results for entropy and support size. Additional estimators for both properties were compared in [OSW16, WY16, WY19, HOSW18,

HO19a] and found to perform similarly to or worse than the estimators we tested, hence we exclude them here. For each property, we considered nine natural-synthetic distributions, shown in Figure 4.1 and 4.2.

Settings We experimented with nine distributions having *support size* $S = 10{,}000$: uniform distribution; a two-steps distribution with probability values $0.5S^{-1}$ and $1.5S^{-1}$; Zipf distribution with power $1/2$; Zipf distribution with power 1; binomial distribution with success probability 0.3; geometric distribution with success probability 0.9; Poisson distribution with mean $0.3S$; a distribution drawn from Dirichlet prior with parameter 1; a distribution drawn from Dirichlet prior with parameter 1/2.

The geometric, Poisson, and Zipf distributions were truncated at S and re-normalized. The horizontal axis shows the number of samples, n, ranging from $S^{0.2}$ to S. Each experiment was repeated 100 times and the reported results, shown on the vertical axis, reflect their mean values and standard deviations. Specifically, the real property value is drawn as a dashed black line, and the other estimators are color/shape coded, with the solid line displaying their mean estimate, and the shaded area corresponding to one standard deviation.

We compared the estimators' performance with n samples to that of two other recent estimators as well as the empirical estimator with n, $n\sqrt{\log A}$, and $n\log A$ samples, where for Shannon entropy, $A = n$ and for support size, $A = S_p$, the actual distribution support size (which is S). We chose the parameter $\varepsilon = 1$. The graphs denote our proposed estimator by Proposed, $\hat{F}^{\mathrm{E}}$ with n samples by Empirical, $\hat{F}^{\mathrm{E}}$ with $n\sqrt{\log A}$ samples by Empirical+, $\hat{F}^{\mathrm{E}}$ with $n\log A$ samples by Empirical++, the entropy and support-size estimators in [WY16] and [WY19] by WY.

Results As Theorem 1 and 4 would imply and the experiments confirmed, for both properties, the proposed estimators with n samples achieved the accuracy as the empirical estimators with at least $n\log n$ samples for entropy and $n\log S_p$ samples for support size. In particular, for entropy, the proposed estimator with n samples performed significantly better than the $n\log n$-sample empirical estimator, for all tested distributions and all values of sample size n.

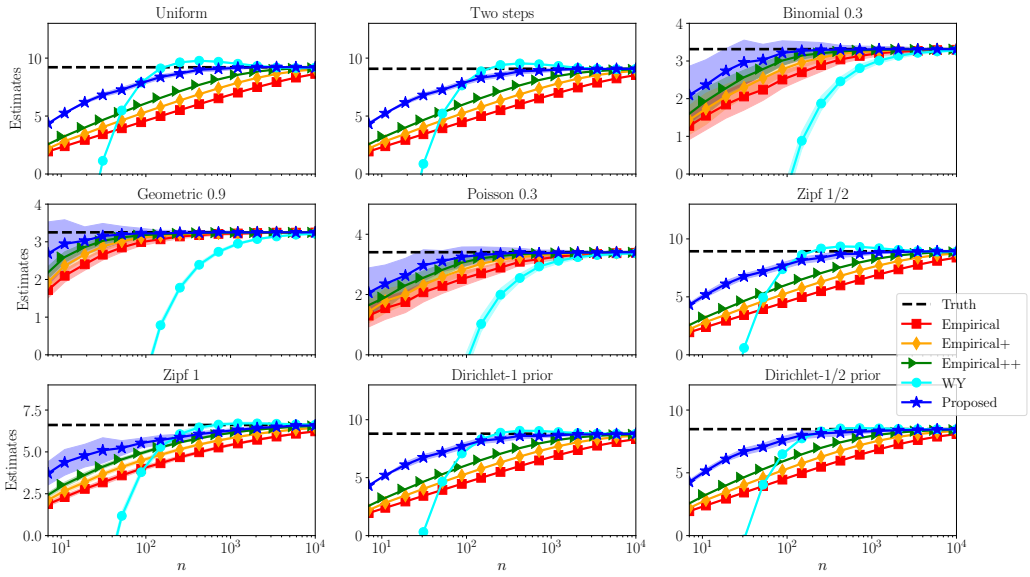


Figure 4.1: Shannon entropy estimation

For both properties, the proposed estimators outperformed the state-of-the-art estimators in terms of accuracy and stability regarding distribution structures.

In Figure 4.1, the horizontal axis is in logarithmic scale. The WY curve is flipped vertically around Truth for all the curves to have similar trends. Besides the samples, the WY estimator takes as input an upper bound of the support size, which is set to be the actual support size in the experiments. The vertical axis shows only nonnegative values.

In Figure 4.2, the horizontal axis is in logarithmic scale. Besides the samples, the WY estimator takes as input a lower bound of the smallest positive probability $p^+_{\min}$, which is set to be $\max\{1/(10S), 4p^+_{\min}\}$ in the experiments. Here, $1/(10S)$ is used to avoid division by zero in numerical computation, and factor 4 represents a reasonable uncertainty about $p^+_{\min}$. For several distributions, such as uniform and geometric, knowing $p^+_{\min}$ yields the full knowledge of the entire probability multiset. Finally, while estimator WY's bias is slightly lower on a few distributions, the corresponding standard deviation is too high to be acceptable.

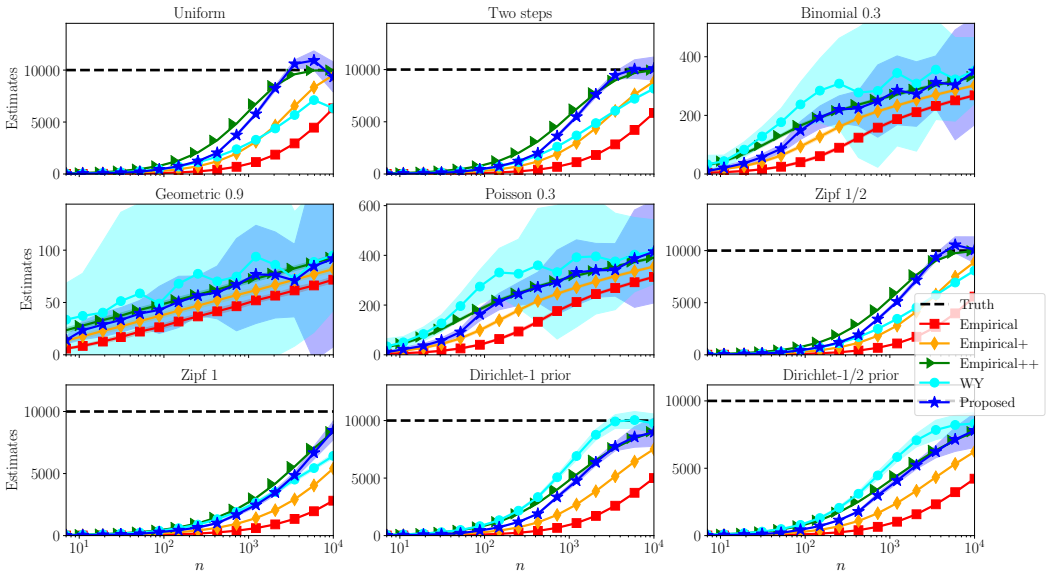


Figure 4.2: Support size estimation

4.6 Acknowledgements

Chapter 4, in full, is a reprint of the material as it appears in *International Conference on Machine Learning*. Yi Hao and Alon Orlitsky, 2020. The dissertation author was the primary investigator and author of this paper.

4.A Appendix Outline

For notational convenience, let $h(p) := -p \log p$ for entropy, $\ell_q(p) := |p-q| - q$ for L_1 distance, $s(p) := \mathbb{1}_{p>0}$ for support size, and $c(p) := 1-(1-p)^m$ for support coverage. Below, we provide an outline of the remaining contents and a high-level overview of our techniques.

In the main body, we focus on Shannon entropy and prove a weaker version of Theorem 1.

Theorem 6. *For all* $\varepsilon \leq 1$ *and all distributions* p*, the estimator* $\hat{H}$ *described in Section 4.D*

satisfies

$$L_{\hat{H}}(p,n) - L_{\hat{H}^E}(p,\varepsilon n\log n) \le (1+c\cdot\varepsilon) \wedge \left(\frac{S_p}{\varepsilon n} + \frac{1}{n^{0.49}}\right).$$

The proof of Theorem 6 in the rest of the paper is organized as follows. In Section 4.B, we present a few useful concentration inequalities for Poisson and binomial random variables. In Section 4.4.1, we relate the n-sample empirical estimator's bias to the degree-n Bernstein polynomial $B_n(h,x)$ via $B_n(h,p_i) = \mathbb{E}[h(N_i/n)]$. In Section 4.C.1, we show that the absolute difference between the *derivative* of $B_n(h,x)$ and a simple function $h_n(x)$ is at most 1, uniformly for all $x \le 1-(n-1)^{-1}$.

Let $a := \varepsilon\log n$ be an amplification parameter. In Section 4.C.2, we approximate $h_{na}(x)$ by a degree-$\Theta(\log n)$ polynomial $\tilde{h}_{na}(x)$ and bound the approximation error uniformly by $c\cdot\varepsilon$. Let $\tilde{H}_{na}(x) := \int_0^x \tilde{h}_{na}(t)dt$. By construction, $|B'_{na}(h,x) - \tilde{h}_{na}(x)| \le |B'_{na}(h,x) - h_{na}(x)| + |h_{na}(x) - \tilde{h}_{na}(x)| \le 1 + c\cdot\varepsilon$, implying $|\tilde{H}_{na}(x) - B_{na}(h,x)| \le x(1+c\cdot\varepsilon)$.

In Section 4.D, we construct our estimator $\hat{H}$ as follows.

First, we divide the symbols into small- and large- probability symbols according to their counts in an independent n-element sample sequence. The concentration inequalities in Section 4.B imply that this step can be performed with relatively high confidence. Then, we estimate the partial entropy of each small-probability symbol i with a near-unbiased estimator of $\tilde{H}_{na}(p_i)$, and the combined partial entropy of the large-probability symbols with a simple variant of the empirical estimator. The final estimator is the sum of these small- and large- probability estimators.

In Section 4.E, we bound the bias of $\hat{H}$. In Sections 4.E.1 and 4.E.2, we use properties of $\tilde{H}_{na}$ and the Bernstein polynomials to bound the partial biases of the small- and large-probability estimators in terms of n, respectively. The critical observation is $|\sum_i(\tilde{H}_{na}(p_i) - B_{na}(h,p_i))| \le \sum_i p_i(1+c\cdot\varepsilon) = 1+c\cdot\varepsilon$, implying that the small-probability estimator has a low bias. To bound the bias of the large-probability estimator, we principally rely on the elegant inequality $|B_n(h,x) - h(x)| \le 1/n$.

By the triangle inequality, it remains to bound the mean absolute deviation of $\hat{H}$. We bound this quantity by bounding the partial variances of the small- and large- probability estimators in Section 4.F.1 and Section 4.F.2, respectively. Intuitively speaking, the small-probability estimator has a small variance because it is constructed based on a low-degree polynomial; the large-probability estimator has a small variance because $h(x)$ is smoother for larger values of x.

To demonstrate the efficacy of our methods, in Section 4.5, we compare the experimental performance of our estimators with that of the state-of-the-art property estimators for Shannon entropy and support size over nine distributions. Our competitive estimators outperformed these existing algorithms on nearly all the experimented instances.

Replacing the simple function $h_n(x)$ by a much finer approximation of $B_n(h,x)$ based on *differential smoothing*, we establish the full version of Theorem 1 in Appendix 4.H. Applying similar techniques, we prove the other four results in Appendices 4.I (Theorem 2 and 3), 4.J (Theorem 4), and 4.K (Theorem 5).

Computational complexity Section 4.G presents the Remez algorithm [Rem34, PT09, Tre13] for computing the best polynomial approximation of a function, and shows that it takes only $\tilde{O}(n)$ time to compute our approximation-based estimators.

4.B Concentration Inequalities

The next lemma gives tight tail probability bounds for Poisson/binomial random variables.

Lemma 6 ([GL17])**.** *Let X be a Poisson or binomial random variable with mean μ, then for any $\delta > 0$,*

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leq e^{-(\delta^2 \wedge \delta)\mu/3},$$

and for any $\delta \in (0,1)$,

$$\mathbb{P}(X \le (1-\delta)\mu) \le \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \le e^{-\delta^2\mu/2}.$$

4.C Approximating Bernstein Polynomials

With n samples, the bias of the empirical estimator in estimating $H(p)$ is

$$\mathrm{Bias}_n(\hat{H}^E, p) := \mathbb{E}[\hat{H}^E(X^n)] - H(p).$$

By the linearity of expectation, the right-hand side equals

$$\mathbb{E}[\hat{H}^E(X^n)] - H(p) = \sum_{i\in[k]} \left(\mathbb{E}\left[h\left(\frac{N_i}{n}\right)\right] - h(p_i)\right).$$

Noting that the degree-n Bernstein polynomial of h is

$$B_n(h,x) := \mathop{\mathbb{E}}_{N_x\sim\mathrm{bin}(n,x)}\left[h\left(\frac{N_x}{n}\right)\right] = \sum_{j=0}^{n} h\left(\frac{j}{n}\right)\binom{n}{j}x^j(1-x)^{n-j},$$

we can express the bias of the empirical estimator as

$$\mathrm{Bias}_n(\hat{H}^E, p) = \sum_{i\in[k]} \left(B_n(h,p_i) - h(p_i)\right).$$

Given a sampling number n and a parameter $\varepsilon \le 1$, define the amplification factor $a := \varepsilon \log n$. Let c_l and c_s be sufficiently large and small absolute constants, respectively. In the following sections, we find a polynomial $\tilde{h}_{na}(x)$ of degree $d-1 := d_n - 1 := c_s \log n - 1$, whose error in approximating $B'_{na}(h,x)$ over $I_n := [0,\tau_n] := [0, c_l(\log n)/n]$ satisfies

$$|B'_{na}(h,x) - \tilde{h}_{na}(x)| \le 1 + O(\varepsilon).$$

By the triangle inequality of integrals, the degree-d polynomial

$$\tilde{H}_{na}(x) := \int_0^x \tilde{h}_{na}(t)dt,$$

approximates $B_{na}(h,x)$ with the following pointwise error guarantee.

Lemma 7. *For any $x \in I_n$,*

$$|B_{na}(h,x) - \tilde{H}_{na}(x)| \leq x\,(1 + O(\varepsilon)).$$

In Section 4.C.1, we relate $B'_n(h,x)$ to a simple function $h_n(x)$ that can be expressed in terms of $h(x)$. In Section 4.C.2, we approximate $h_n(x)$ by a linear combination of degree-d min-max polynomials of $h(x)$ over different intervals. The resulting polynomial is $\tilde{h}_{na}(x)$.

4.C.1 Derivative of Bernstein Polynomials

According to [Bus17], the first-order derivative of the Bernstein polynomial $B_n(h,x)$ is

$$B'_n(h,x) := \sum_{j=0}^{n-1} n\left(h\left(\frac{j+1}{n}\right) - h\left(\frac{j}{n}\right)\right)\binom{n-1}{j}x^j(1-x)^{(n-1)-j}.$$

Hence, letting

$$h_n(x) := n\left(h\left(\left(\frac{n-1}{n}\right)x + \frac{1}{n}\right) - h\left(\left(\frac{n-1}{n}\right)x\right)\right),$$

we can write derivative B'_n as

$$B'_n(h,x) = \sum_{j=0}^{n-1} h_n\left(\frac{j}{n-1}\right)\binom{n-1}{j}x^j(1-x)^{(n-1)-j} = B_{n-1}(h_n,x).$$

Recall that $h(x) = -x\log x$. After some algebra, we get

$$h_n(x) = -\log\left(\frac{n-1}{n}\right) + (n-1)\left(h\left(x+\frac{1}{n-1}\right) - h(x)\right).$$

Furthermore, utilizing analytical attributes of $h(x)$ [BLM72], we can bound the absolute difference between $h(x)$ and its Bernstein polynomial as follows.

Lemma 8. *For any $m > 0$ and $x \in [0,1]$,*

$$-\frac{1-x}{m} \leq B_m(h,x) - h(x) \leq 0.$$

As an immediate corollary,

Corollary 2. *For any $x \in [0, 1-(n-1)^{-1}]$,*

$$|B_n'(h,x) - h_n(x)| = |B_{n-1}(h_n,x) - h_n(x)| \leq 1.$$

Proof. Given the equality $B_n'(h,x) = B_{n-1}(h_n,x)$ for $x \in [0,1-(n-1)^{-1}]$,

$$\begin{aligned}
|B_{n-1}(h_n,x) - h_n(x)| &\leq (n-1)|(B_{n-1}(h,x+(n-1)^{-1}) - h(x+(n-1)^{-1})) \\
&\quad - (B_{n-1}(h,x) - h(x))| \\
&\leq (n-1)\left|\max\left\{\frac{1-x-(n-1)^{-1}}{n-1}, \frac{1-x}{n-1}\right\}\right| \\
&\leq 1,
\end{aligned}$$

where the second inequality follows by Lemma 8. □

4.C.2 Approximating the Derivative Function

Denote the degree-d min-max polynomial of h over $[0,1]$ by

$$\tilde{h}(x) := \sum_{j=0}^{d} b_j x^j.$$

As shown in [WY16], the coefficients of $\tilde{h}(x)$ satisfy

$$|b_j| \lesssim 2^{3d},$$

and the error of $\tilde{h}(x)$ in approximating $h(x)$ admits

$$\max_{x\in[0,1]} |h(x) - \tilde{h}(x)| \lesssim \frac{1}{\log^2 n}.$$

By a change of variables, the degree-d min-max polynomial of h over $I_n = [0, c_l \log n/n]$ is

$$\tilde{h}_1(x) := \sum_{j=0}^{d} b_j \left(\frac{n}{c_l \log n}\right)^{j-1} x^j + \left(\log \frac{n}{c_l \log n}\right) x.$$

Correspondingly, for any $x \in I_n$, we have

$$\max_{x\in I_n} |h(x) - \tilde{h}_1(x)| \lesssim \frac{1}{n \log n}.$$

To approximate $h_{na}(x)$, we approximate $h(x)$ by $\tilde{h}_1(x)$, and $h(x+(na-1)^{-1})$ by $\tilde{h}_1(x+(na-1)^{-1})$. Then, the resulting polynomial is

$$\begin{aligned}
\tilde{h}_{na}(x) &:= -\log \frac{na-1}{na} + (na-1)\left(\tilde{h}_1(x+(na-1)^{-1}) - \tilde{h}_1(x)\right) \\
&= -\log \frac{na-1}{c_l a \log n} + (na-1)\left(\sum_{j=0}^{d} b_j \left(\frac{n}{c_l \log n}\right)^{j-1} \left(\left(x + \frac{1}{na-1}\right)^j - x^j\right)\right).
\end{aligned}$$

By the above reasoning, the error of $\tilde{h}_{na}$ in approximating h_{na} over I_n satisfies

$$\max_{x\in I_n} |h_{na}(x) - \tilde{h}_{na}(x)| \lesssim \frac{na}{n\log n} \lesssim \varepsilon.$$

Moreover, by an application of Corollary 2,

$$\max_{x\in[0,1/2]} |B'_{na}(h,x) - h_{na}(x)| = \max_{x\in[0,1/2]} |B_{na-1}(h_{na},x) - h_{na}(x)| \le 1.$$

The triangle inequality combines the above two inequalities and yields

$$\max_{x\in I_n} |B'_{na}(h,x) - \tilde{h}_{na}(x)| \le 1 + O(\varepsilon).$$

Therefore, denoting

$$\tilde{H}_{na}(x) := \int_0^x \tilde{h}_{na}(t)dt,$$

and noting that $B_{na}(h,0) = 0$, we have

Lemma 9. *For any $x \in I_n$,*

$$|B_{na}(h,x) - \tilde{H}_{na}(x)| \le \int_0^x |B'_{na}(h,t) - \tilde{h}_{na}(t)|dt \le x(1 + O(\varepsilon)).$$

4.D A Competitive Entropy Estimator

In this section, we design an explicit entropy estimator $\hat{H}$ based on $\tilde{H}_{na}$ and the empirical estimator. Note that $\tilde{H}_{na}(x)$ is a polynomial with a zero constant term. For $t \ge 1$, denote

$$g_t := \sum_{j=t}^{d} \frac{b_j}{j+1} \left(\frac{n}{c_l \log n}\right)^{j-1} \left(\frac{1}{na-1}\right)^{j-t} \binom{j+1}{j-t+1}.$$

Setting $b'_t = g_t$ for $t \ge 2$ and $b'_1 = g_1 - \log\frac{na-1}{c_l a \log n}$, we have the following lemma.

Lemma 10. *The function $\tilde{H}_{na}(x)$ can be written as*

$$\tilde{H}_{na}(x) = \sum_{t=1}^{d} b'_t x^t.$$

In addition, its coefficients satisfy

$$|b'_t| \lesssim 2^{4d} \left(\frac{n}{c_l \log n} \right)^{t-1}.$$

The proof of the above lemma is postponed to the end of this section.

To simplify our analysis and remove the dependency between symbol counts, we use the conventional *Poisson sampling* technique [WY16, ADOS17]. Specifically, instead of drawing exactly n samples, we make the sample size an independent Poisson random variable N with mean n. This does not change the statistical nature of the problem as $N \sim \mathrm{Poi}(n)$ highly concentrates around its mean (see Lemma 6). We still define N_i as the count of symbol i in X^N. Due to Poisson sampling, these counts are now mutually independent and satisfy $N_i \sim \mathrm{Poi}(np_i),\ \forall i \in [k]$.

For each $i \in [k]$, let $N_i^{\underline{t}} := \prod_{m=0}^{t-1}(N_i - m)$ be the t-th order *falling factorial* of N_i. The following identity is well-known:

$$\mathbb{E}[N_i^{\underline{t}}] = (np_i)^t,\ \forall t \le n.$$

Note that for sufficiently small c_s or sufficiently large n, the degree parameter $d = c_s \log n \le n, \forall n$. By the linearity of expectation, the unbiased estimator of $\tilde{H}_{na}(p_i)$ is

$$\hat{H}_{na}(N_i) := \sum_{t=1}^{d} b'_t \frac{N_i^{\underline{t}}}{n^t}.$$

Let N' be an independent Poisson variable with mean n, and $X^{N'}$ be an independent length-N' sample sequence drawn from p. Analogously, we denote by N'_i the number of times that symbol

$i \in [k]$ appears. Depending on whether $N_i' > \varepsilon^{-1}$ or not, we classify $p_i, i \in [k]$, into two categories: small- and large- probabilities. For small probabilities, we apply a simple variant of $\hat{H}_{na}(N_i)$; for large probabilities, we estimate $h(p_i)$ by an empirical-estimator variant.

Specifically, for each $i \in [k]$, we estimate $h(p_i)$ by

$$\hat{h}(N_i, N_i') := \hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \le c_l \log n} \cdot \mathbb{1}_{N_i' \le \varepsilon^{-1}} + h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N_i' > \varepsilon^{-1}}.$$

Consequently, we approximate $H(p)$ by

$$\hat{H}(X^N, X^{N'}) := \sum_{i \in [k]} \hat{h}(N_i, N_i').$$

For the simplicity of illustration, we will refer to

$$\hat{H}_S(X^N, X^{N'}) := \sum_{i \in [k]} \hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \le c_l \log n} \cdot \mathbb{1}_{N_i' \le \varepsilon^{-1}}$$

as the *small-probability estimator*, and

$$\hat{H}_L(X^N, X^{N'}) := \sum_{i \in [k]} h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N_i' > \varepsilon^{-1}}$$

as the *large-probability estimator*. Then, $\hat{H}$ is the sum of these two estimators.

In the next two sections, we analyze the bias and mean absolute deviation of $\hat{H}$. In Section 4.E, we show that for any p, the absolute bias of $\hat{H}$ satisfies

$$\left|\mathbb{E}[\hat{H}(X^N, X^{N'})] - H(p)\right| \le \left|\mathrm{Bias}(\hat{H}^E, na)\right| + (1 + O(\varepsilon)) \left(1 \wedge (\varepsilon^{-1} + 1)\frac{S_p}{n}\right).$$

In Section 4.F, we further show that the mean absolute deviation of $\hat{H}$ satisfies

$$\mathbb{E}\left|\hat{H}(X^N, X^{N'}) - \mathbb{E}[\hat{H}(X^N, X^{N'})]\right| \lesssim \frac{1}{n^{1-\Theta(c_s)}}.$$

For sufficiently small c_s, the triangle inequality combines the above inequalities and yields

$$\mathbb{E}\left|\hat{H}(X^N, X^{N'}) - H(p)\right| \leq \mathrm{Bias}(\hat{H}^E, na) + (1 + c \cdot \varepsilon) \wedge \left(\frac{S_p}{\varepsilon n} + \frac{1}{n^{0.49}}\right).$$

This basically completes the proof of Theorem 6.

Proof of Lemma 10

We begin by proving the first claim:

$$\tilde{H}_{na}(x) = -\sum_{t=1}^{d} b'_t x^t.$$

By definition, $\tilde{H}_{na}(x)$ satisfies

$$\begin{aligned}
&\tilde{H}_{na}(x) + \left(\log \frac{na-1}{c_l a \log n}\right) x \\
&= (na-1)\left(\sum_{j=1}^{d} \frac{b_j}{j+1}\left(\frac{n}{c_l \log n}\right)^{j-1}\left(\left(x + \frac{1}{na-1}\right)^{j+1} - \left(\frac{1}{na-1}\right)^{j+1} - x^{j+1}\right)\right) \\
&= \sum_{j=1}^{d} \frac{b_j}{j+1}\left(\frac{n}{c_l \log n}\right)^{j-1}\left(\sum_{m=0}^{j-1}\left(\frac{1}{na-1}\right)^{m} x^{j-m}\binom{j+1}{m+1}\right) \\
&= \sum_{t=1}^{d} x^t \left(\sum_{j=t}^{d} \frac{b_j}{j+1}\left(\frac{n}{c_l \log n}\right)^{j-1}\left(\frac{1}{na-1}\right)^{j-t}\binom{j+1}{j-t+1}\right),
\end{aligned}$$

where the last step follows by reorganizing the indices.

Next we establish the second claim. Recall that $d = c_s \log n$, thus,

$$\log \frac{na-1}{c_l a \log n} \lesssim 2^{4d}.$$

Since $b'_t = g_t$ for $t \geq 2$ and $b'_1 = g_1 - \log \frac{na-1}{c_l a \log n}$, it suffices to bound the magnitude of g_t:

$$\begin{aligned}
|g_t| &\leq \sum_{j=t}^{d} \frac{|b_j|}{j+1} \left(\frac{n}{c_l \log n}\right)^{j-1} \left(\frac{1}{na-1}\right)^{j-t} \binom{j+1}{j-t+1} \\
&\leq \sum_{j=t}^{d} |b_j| \left(\frac{1}{c_l \log n}\right)^{j-1} n^{t-1} \binom{j}{t} \\
&\leq \left(\frac{n}{c_l \log n}\right)^{t-1} \sum_{j=t}^{d} |b_j| \binom{j}{t} \\
&\leq \left(\frac{n}{c_l \log n}\right)^{t-1} \sum_{j=t}^{d} |b_j| \binom{d}{j-t} \\
&\lesssim 2^{4d} \left(\frac{n}{c_l \log n}\right)^{t-1}.
\end{aligned}$$

4.E Bounding the Bias of $\hat{H}$

By the triangle inequality, the absolute bias of $\hat{H}$ in estimating $H(p)$ satisfies

$$\begin{aligned}
\left| \sum_{i \in [k]} (\mathbb{E}[\hat{h}(N_i, N'_i)] - h(p_i)) \right| &\leq \left| \sum_{i \in [k]} (B_{na}(h, p_i) - h(p_i)) \right| \\
&\quad + \left| \sum_{i \in [k]} (\mathbb{E}[\hat{h}(N_i, N'_i)] - B_{na}(h, p_i)) \right|.
\end{aligned}$$

Note that the first term on the right-hand side is the absolute bias of the empirical estimator with sample size $na = \varepsilon n \log n$, that is,

$$\mathrm{Bias}_{na}(\hat{H}^E, p) = \left| \sum_{i \in [k]} (B_{na}(h, p_i) - h(p_i)) \right|.$$

Hence, we need to consider only the second term on the right-hand side, which admits

$$\left| \sum_{i \in [k]} (\mathbb{E}[\hat{h}(N_i, N_i')] - B_{na}(h, p_i)) \right| \leq \text{Bias}_S + \text{Bias}_L,$$

where

$$\text{Bias}_S := \left| \sum_{i \in [k]} \mathbb{E}\left[\left(\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \leq c_l \log n} - B_{na}(h, p_i) \right) \cdot \mathbb{1}_{N_i' \leq \varepsilon^{-1}} \right] \right|$$

is the absolute bias of the small-probability estimator, and

$$\text{Bias}_L := \left| \sum_{i \in [k]} \mathbb{E}\left[\left(h\left(\frac{N_i}{n}\right) - B_{na}(h, p_i) \right) \cdot \mathbb{1}_{N_i' > \varepsilon^{-1}} \right] \right|$$

is the absolute bias of the large-probability estimator.

Assume that c_l is sufficiently large. In Section 4.E.1, we bound the small-probability bias by

$$\text{Bias}_S \leq (1 + O(\varepsilon)) \left(1 \wedge (\varepsilon^{-1} + 1) \frac{S_p}{n} \right).$$

In Section 4.E.2, we bound the large-probability bias by

$$\text{Bias}_L \leq 2 \left(\varepsilon \wedge \frac{S_p}{n} \right).$$

4.E.1 Bias of the Small-Probability Estimator

We first consider and analyze Bias_S. By the triangle inequality,

$$\begin{aligned}
\text{Bias}_S \leq & \sum_{i: p_i \notin I_n} \left| \mathbb{E}[\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \leq c_l \log n}] - B_{na}(h, p_i) \right| \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \\
& + \sum_{i: p_i \in I_n} \left| \mathbb{E}\left[\hat{H}_{na}(N_i) \right] - B_{na}(h, p_i) \right| \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \\
& + \sum_{i: p_i \in I_n} \left| \mathbb{E}\left[\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i > c_l \log n} \right] \cdot \mathbb{E}\left[\mathbb{1}_{N_i' \leq \varepsilon^{-1}} \right] \right|.
\end{aligned}$$

Let us assume $\varepsilon \log n \geq 1$ and consider the first sum on the right-hand side. By the general reasoning in the proof of Lemma 11, we can show that

$$\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \leq c_l \log n} \lesssim 2^{5d} \cdot \frac{\log^2 n}{n}.$$

Further assume that c_s and c_l are sufficiently small and large, respectively. For large enough n, the above inequality bounds the first sum by

$$\sum_{i: p_i \notin I_n} \left| \hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \leq c_l \log n} - B_{na}(h, p_i) \right| \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \leq \sum_{i: p_i \notin I_n} \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \leq \frac{1}{n^5} \cdot \frac{n}{c_l \log n} \leq \frac{1}{n^4}.$$

For the second sum on the right-hand side, by Lemma 9,

$$\begin{aligned}
\sum_{i: p_i \in I_n} \left| \mathbb{E}\left[\hat{H}_{na}(N_i)\right] - B_{na}(h, p_i) \right| \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] &\leq \sum_{i: p_i \in I_n} \left| \mathbb{E}\left[\hat{H}_{na}(N_i)\right] - B_{na}(h, p_i) \right| \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \\
&= \sum_{i: p_i \in I_n} \left| \tilde{H}_{na}(p_i) - B_{na}(h, p_i) \right| \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \\
&\leq \sum_{i: p_i \in I_n} (1 + O(\varepsilon))\, p_i \cdot \mathbb{E}[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}] \\
&\leq (1 + O(\varepsilon)) \left(1 \wedge (\varepsilon^{-1} + 1) \frac{S_p}{n} \right).
\end{aligned}$$

The following lemma bounds the last sum and completes our argument.

Lemma 11. *For sufficiently large c_l,*

$$\sum_{i \in [k]} \left| \mathbb{E}\left[\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i > c_l \log n} \right] \cdot \mathbb{E}\left[\mathbb{1}_{N_i' \leq \varepsilon^{-1}} \right] \right| \leq \frac{1}{n^5}.$$

Proof. For simplicity, we assume that $c_l \geq 4$ and $\varepsilon \log n \geq 1$. By the triangle inequality,

$$\begin{aligned}
&\left| \mathbb{E}\left[\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i > c_l \log n}\right] \cdot \mathbb{E}\left[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}\right] \right| \\
&\leq \sum_{j=1}^{\infty} \left| \mathbb{E}\left[\hat{H}_{na}(N_i) \cdot \mathbb{1}_{c_l(j+1)\log n \geq N_i > c_l j \log n}\right] \cdot \mathbb{E}\left[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}\right] \right|.
\end{aligned}$$

To bound the last term, we rely on the following result: For any $j \geq 1$,

$$\left| \mathbb{E}\left[\mathbb{1}_{c_l(j+1)\log n \geq N_i > c_l j \log n}\right] \cdot \mathbb{E}\left[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}\right] \right| \leq \left(1 + \varepsilon^{-1}\right) np_i \cdot e^{-\Theta(c_l j \log n)}.$$

To prove this inequality, we apply Lemma 6 and consider two cases:

Case 1: If $np_i < (3c_l/8) j \log n$, then

$$\mathbb{E}\left[\mathbb{1}_{c_l(j+1)\log n \geq N_i > c_l j \log n}\right] \leq np_i \cdot e^{-\Theta(c_l j \log n)}.$$

Case 2: If $np_i \geq (3c_l/8) j \log n$, then

$$\mathbb{E}\left[\mathbb{1}_{N_i' \leq \varepsilon^{-1}}\right] \leq np_i \varepsilon^{-1} \cdot e^{-\Theta(c_l j \log n)}.$$

This essentially completes the proof. Next, we bound $\hat{H}_{na}(N_i)$ for $N_i \in [c_l j \log n, c_l(j+1)\log n]$:

$$\begin{aligned}
|\hat{H}_{na}(N_i)| &= \left| \left(\log \frac{na-1}{c_l a \log n} \right) \frac{N_i}{n} + \sum_{t=1}^{d} b_t' \frac{N_i^{\underline{t}}}{n^t} \right| \\
&\lesssim 2^{4d} \cdot \sum_{t=1}^{c_s \log n} \left(\frac{n}{c_l \log n} \right)^{t-1} \frac{(c_l(j+1)\log n)^t}{n^t} \\
&\lesssim 2^{5d} \cdot \frac{c_l j \log n}{n} \sum_{t=1}^{c_s \log n} j^{t-1} \\
&\lesssim 2^{5d} \cdot \frac{c_l j \log n}{n} (j^{c_s \log n} + c_s \log n).
\end{aligned}$$

Hence, for sufficiently large c_l,

$$\begin{aligned}
&\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i>c_l\log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N_i'\le \varepsilon^{-1}}\right]\right| \\
&\le \sum_{j=1}^{\infty}\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot \mathbb{1}_{c_l(j+1)\log n\ge N_i>c_l j\log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N_i'\le \varepsilon^{-1}}\right]\right| \\
&\le \sum_{j=1}^{\infty} O(2^{5d})\cdot c_l j\log n(j^{c_s\log n}+c_s\log n)\cdot \mathbb{E}\left[\mathbb{1}_{c_l(j+1)\log n\ge N_i>c_l j\log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N_i'\le \varepsilon^{-1}}\right] \\
&\lesssim 2^{5d}\cdot \sum_{j=1}^{\infty}\left(1+\varepsilon^{-1}\right)p_i\cdot e^{-\Theta(c_l j\log n)}\cdot c_l j\log n(j^{c_s\log n}+c_s\log n) \\
&\le p_i\sum_{j=1}^{\infty}\frac{1}{2n^{5j}} \\
&\le \frac{p_i}{n^5}.
\end{aligned}$$

Summing the right-hand side over $i\in[k]$ yields the desired result. □

4.E.2 Bias of the Large-Probability Estimator

This section proves the bias bound $\mathrm{Bias}_L \le 2\left(\varepsilon \wedge (S_p/n)\right)$. By the triangle inequality,

$$\begin{aligned}
\mathrm{Bias}_L &\le \sum_{i\in[k]}\left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right)-B_{na}(h,p_i)\right]\right|\cdot \mathbb{E}\left[\mathbb{1}_{N_i'>\varepsilon^{-1}}\right] \\
&\le \sum_{i\in[k]}\left|h(p_i)-B_{na}(h,p_i)\right|\cdot \mathbb{E}\left[\mathbb{1}_{N_i'>\varepsilon^{-1}}\right]+\sum_{i\in[k]}\left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right)-h(p_i)\right]\right|\cdot \mathbb{E}\left[\mathbb{1}_{N_i'>\varepsilon^{-1}}\right].
\end{aligned}$$

We need the following inequality to bound the right-hand side.

$$0\le x\log x-(x-1)\le (x-1)^2,\ \forall x\in[0,1].$$

For simplicity, denote $\hat{p}_i := N_i/n$. Then,

$$\begin{aligned}
\left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right) - h(p_i)\right]\right| &= |\mathbb{E}[p_i \log p_i - \hat{p}_i \log \hat{p}_i]| \\
&\leq |\mathbb{E}[p_i \log p_i - \hat{p}_i \log p_i]| + |\mathbb{E}[\hat{p}_i \log p_i - \hat{p}_i \log \hat{p}_i]| \\
&= p_i \cdot \left|\mathbb{E}\left[\frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i}\right]\right| \\
&\leq p_i \cdot \left|\mathbb{E}\left[\left(\frac{\hat{p}_i}{p_i} - 1\right) + \left(\frac{\hat{p}_i}{p_i} - 1\right)^2\right]\right| \\
&= \frac{1}{n}.
\end{aligned}$$

Replacing n by na in the above argument yields

$$|h(p_i) - B_{na}(h, p_i)| \leq \frac{1}{na}.$$

Consider the first term on the right-hand side. By the last bound and Markov's inequality,

$$\begin{aligned}
\sum_{i \in [k]} |h(p_i) - B_{na}(h, p_i)| \cdot \mathbb{E}\left[\mathbb{1}_{N_i' > \varepsilon^{-1}}\right] &\leq \frac{1}{na} \sum_{i \in [k]} \mathbb{E}\left[\mathbb{1}_{N_i' > \varepsilon^{-1}}\right] \\
&\leq \frac{1}{na} \sum_{i \in [k]} (\mathbb{1}_{p_i > 0} \wedge \varepsilon n p_i) \\
&\leq \varepsilon \wedge \frac{S_p}{n}.
\end{aligned}$$

For the second term, an analogous argument yields

$$\sum_{i \in [k]} \left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right) - h(p_i)\right]\right| \cdot \mathbb{E}\left[\mathbb{1}_{N_i' > \varepsilon}\right] \leq \varepsilon \wedge \frac{S_p}{n}.$$

4.F Bounding the Mean Absolute Deviation of $\hat{H}$

By Jensen's inequality,

$$\mathbb{E}|\hat{H}(X^N, X^{N'}) - \mathbb{E}[\hat{H}(X^N, X^{N'})]| \leq \sqrt{\mathrm{Var}(\hat{H}(X^N, X^{N'}))}.$$

Hence, to bound the mean absolute deviation of $\hat{H}$, it suffices to bound its variance. Note that the symbol counts are mutually independent. The inequality $\mathrm{Var}(X+Y) \leq 2(\mathrm{Var}(X) + \mathrm{Var}(Y))$ implies

$$\mathrm{Var}(\hat{H}(X^N, X^{N'})) = \sum_{i \in [k]} \mathrm{Var}(\hat{h}(N_i, N_i')) \leq 2\mathrm{Var}_S + 2\mathrm{Var}_L,$$

where

$$\mathrm{Var}_S := \sum_{i \in [k]} \mathrm{Var}\left(\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \leq c_l \log n} \cdot \mathbb{1}_{N_i' \leq \varepsilon^{-1}}\right)$$

is the variance of the small-probability estimator, and

$$\mathrm{Var}_L := \sum_{i \in [k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N_i' > \varepsilon^{-1}}\right)$$

is the variance of the large-probability estimator. Assume that c_l and c_s are sufficiently large and small absolute constants. In Section 4.F.1 and 4.F.2, we will respectively establish

$$\mathrm{Var}_S \lesssim \frac{1}{n^{1-\Theta(c_s)}} \text{ and } \mathrm{Var}_L \lesssim \frac{(\log n)^3}{n}.$$

4.F.1 Variance of the Small-Probability Estimator

First we bound the small-probability variance Var_S and prove $\mathrm{Var}_S \le O\left(1/n^{1-\Theta(c_s)}\right)$. Following the sequence of derivations in Section 4.E.1,

$$\begin{aligned}
\mathrm{Var}_S &\le 2\sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i>c_l\log n}\cdot \mathbb{1}_{N_i'\le \varepsilon^{-1}}\right)\\
&\quad + 2\sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i'\le \varepsilon^{-1}}\right)\\
&\le 2\sum_{i\in[k]} \mathbb{E}[(\hat{H}_{na}(N_i))^2\cdot \mathbb{1}_{N_i>c_l\log n}]\cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}]\\
&\quad + 2\sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i)\right)\cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] + 2\sum_{i\in[k]} (\mathbb{E}[\hat{H}_{na}(N_i)])^2\cdot \mathrm{Var}(\mathbb{1}_{N_i'\le \varepsilon^{-1}})\\
&\le 2\sum_{i\in[k]} \mathbb{E}[(\hat{H}_{na}(N_i))^2\cdot \mathbb{1}_{N_i>c_l\log n}]\cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}]\\
&\quad + 2\sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i)\right)\cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] + 2\sum_{i\in[k]} (\tilde{H}_{na}(p_i))^2\cdot \mathrm{Var}(\mathbb{1}_{N_i'\le \varepsilon^{-1}}),
\end{aligned}$$

where the first step follows by $\mathrm{Var}(X-Y)\le 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$, the second step follows from $\mathrm{Var}(A\cdot B) = \mathbb{E}[A^2]\mathrm{Var}(B)+\mathrm{Var}(A)(\mathbb{E}[B])^2$ for any independent random variables A and B, and the last step follows from our construction, which satisfies $\mathbb{E}[\hat{H}_{na}(N_i)] = \tilde{H}_{na}(p_i)$.

Similar to the proof of Lemma 11, for the first term on the right-hand side and sufficiently large c_l,

$$\sum_{i\in[k]} \left|\mathbb{E}\left[(\hat{H}_{na}(N_i))^2\cdot \mathbb{1}_{N_i>c_l\log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N_i'\le \varepsilon^{-1}}\right]\right| \le \sum_{i\in[k]} \frac{p_i}{n^3} = \frac{1}{n^3}.$$

As for the second term on the right-hand side,

$$
\begin{aligned}
\mathrm{RHS}_2 &:= \sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}\right) \cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] \\
&\lesssim 2^{8d} \cdot \sum_{i\in[k]} d \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} \frac{\mathrm{Var}(N_i^t)}{n^{2t}} \cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} \mathrm{Var}(N_i^t) \cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} (np_i)^t \sum_{j=0}^{t-1} \binom{t}{j} (np_i)^j \frac{t!}{j!} \cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} (np_i)^t (t+np_i)^t \cdot \mathbb{E}[\mathbb{1}_{N_i'\le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} 2^t \left((np_i)^{2t} + (np_i)^t t^t\right) \cdot \Pr(N_i' \le \varepsilon^{-1}) \\
&\le 2^{8d} \cdot \frac{d}{n} \sum_{i\in[k]} p_i \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} 2^t \Big((\varepsilon^{-1}+2t)^{2t-1} \cdot \Pr(N_i' \le \varepsilon^{-1}+2t) \\
&+ (\varepsilon^{-1}+t)^{t-1} t^t \cdot \Pr(N_i' \le \varepsilon^{-1}+t)\Big) \\
&\lesssim 2^{9d} \cdot \frac{d}{n}.
\end{aligned}
$$

It remains to bound the third term. Leveraging $|\tilde{H}_{na}(p_i)| \lesssim p_i 2^{5d}$ shows that

$$
\begin{aligned}
&\sum_{i\in[k]} (\tilde{H}_{na}(p_i))^2 \cdot \mathrm{Var}(\mathbb{1}_{N_i'\le \varepsilon^{-1}}) \\
&\lesssim 2^{8d} \cdot \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} p_i^{2t} \cdot \mathrm{Var}(\mathbb{1}_{N_i'\le \varepsilon^{-1}}) \\
&\le 2^{8d} \cdot \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} p_i^{2t} \cdot \Pr(N_i' \le \varepsilon^{-1}) \\
&= 2^{8d} \cdot \sum_{i\in[k]} p_i \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} p_i^{2t-1} \cdot \sum_{m=0}^{\varepsilon^{-1}} e^{-np_i} \frac{(np_i)^m}{m!} \\
&\le 2^{8d} \cdot \sum_{i\in[k]} p_i \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} \left(\frac{2t-1+\varepsilon^{-1}}{n}\right)^{2t-1} \Pr(N_i \le 2t-1+\varepsilon^{-1}) \\
&\le 2^{8d} \cdot \sum_{i\in[k]} p_i \cdot c_s \log n \cdot \frac{c_l \log n}{n} \\
&\lesssim \frac{2^{9d}}{n}.
\end{aligned}
$$

Consolidating all the three bounds above yields

$$
\mathrm{Var}_S \le \frac{2}{n^3} + O(2^{9d}) \cdot \frac{d}{n} + O\left(\frac{2^{9d}}{n}\right) \le \frac{1}{n^{1-\Theta(c_s)}},
$$

where the last step follows by $d = c_s \log n$.

4.F.2 Variance of the Large-Probability Estimator

In this section we bound the quantity Var_L and establish $\mathrm{Var}_L \lesssim (\log n)^3/n$. Due to independence,

$$
\mathrm{Var}_L = \sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N_i'>\varepsilon^{-1}}\right).
$$

The following lemma bounds the right-hand-side summation.

Lemma 12. *For any integer $s \geq 1$,*

$$\sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N_i'>s}\right) \leq (\log n)^2 \frac{4s}{n}.$$

Proof. First, we effectively decompose the variances:

$$\begin{aligned}\sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\mathbb{1}_{N_i'>s}\right) &= \mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[h^2\left(\frac{N_i}{n}\right)\right] + \sum_{i\in[k]}\left(\mathbb{E}[\mathbb{1}_{N_i'>s}]\right)^2 \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\right)\\ &\leq \mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[h^2\left(\frac{N_i}{n}\right)\right] + \sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\right).\end{aligned}$$

To bound the first term on the right-hand side, note that

$$\begin{aligned}\mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[h^2\left(\frac{N_i}{n}\right)\right] &\leq \mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[(\log n)^2\left(\frac{N_i}{n}\right)^2\right]\\ &\leq (\log n)^2\frac{p_i}{n}\left(1+np_i\mathrm{Var}(\mathbb{1}_{N_i'>s})\right),\end{aligned}$$

where the term in the parentheses further admits

$$\begin{aligned}p_i\,\mathrm{Var}(\mathbb{1}_{N_i'>s}) &\leq p_i\cdot\mathbb{P}[N_i'\leq s]\\ &= e^{-np_i}\sum_{j=0}^{s}\frac{(np_i)^{j+1}}{(j+1)!}\frac{j+1}{n}\\ &\leq \frac{s+1}{n}e^{-np_i}\sum_{j=0}^{s}\frac{(np_i)^{j+1}}{(j+1)!}\\ &= \frac{s+1}{n}\mathbb{P}(1\leq N_x'\leq s+1)\\ &\leq \frac{s+1}{n}.\end{aligned}$$

To bound the second term, let $\hat{N}_i$ be an i.i.d. copy of N_i for each i,

$$\begin{aligned}
2\mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\right) &= \mathrm{Var}\left(h\left(\frac{N_i}{n}\right) - h\left(\frac{\hat{N}_i}{n}\right)\right) \\
&= \mathbb{E}\left(h\left(\frac{N_i}{n}\right) - h\left(\frac{\hat{N}_i}{n}\right)\right)^2 \\
&\leq (\log n)^2 \mathbb{E}\left(\frac{N_i}{n} - \frac{\hat{N}_i}{n}\right)^2 \\
&= 2(\log n)^2 \cdot \frac{p_i}{n}.
\end{aligned}$$

A simple combination of these bounds yields the lemma. □

Setting $s = \varepsilon^{-1}$ in Lemma 12 and assuming $\varepsilon \log n \geq 1$, we obtain

$$\mathrm{Var}_L = \sum_{i \in [k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N_i' > \varepsilon^{-1}}\right) \leq \frac{4(\log n)^3}{n}.$$

4.G Computational Complexity

The dominant computation step is finding the min-max polynomial of $B_m'(h,x)$, in which we use the well-known Remez algorithm [PT09, Tre13]. Below, we shall argue that the algorithm takes only $\tilde{O}(n)$ time (*number of bit operations*) to well approximate $B_m'(h,x)$.

4.G.1 Remez Algorithm

The algorithm named after [Rem34] is an efficient iterative algorithm that numerically computes the minimax polynomial. For a valid domain $[a,b]$, set our objective to well approximating the function $f(x) : [a,b] \to \mathbb{R}$ by a degree-d real polynomial $P(x)$, in the min-max sense. We briefly illustrate a simple version of the algorithm below.

1. There are several different ways to initialize the algorithm. A popular initialization is to

use the *Chebyshev nodes*. Specifically, we compute $d+2$ points $x_0, x_1, \ldots x_{d+1}$ as

$$x_i := \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\cos\left(\frac{2i+1}{2(d+2)}\pi\right), i = 0, 1, \ldots, d+1.$$

2. For $x_0, x_1, \ldots x_{d+1}$, solve the linear system of $d+2$ equations

$$b_0 + b_1 \cdot x_i + \ldots + b_d \cdot x_i^d + E \cdot (-1)^i = f(x_i) \ \ (\text{where } i = 0, 1, \ldots d+1),$$

for the unknowns $b_0, b_1, \ldots b_d,$ and E.

3. (Re)form the polynomial $P(x)$ as

$$P(x) := b_0 + b_1 \cdot x + \ldots + b_d \cdot x^d.$$

4. Compute the $d+2$ local extrema of the error function

$$\mathcal{E}(x) := P(x) - f(x)$$

over the sign-invariant regions, and denote them by $x_0^*, \ldots x_{d+1}^*$, sorted in descending order.

5. Replace x_i by x_i^* for $i = 0, 1, \ldots d+1$ and go back to Step 2 until quantity E converges.

Next, we analyze the time complexity of the Remez algorithm when applied to our setting.

4.G.2 Complexity of Evaluating $f(x)$

To compute our estimator, the function to approximate is the degree-$\tilde{\Theta}(n)$ polynomial $f(x) := B_m(h_m, \tau_n \cdot x)$ with $m = na - 1$ (different from the prior version to simply the notation), $a \in [1, \log n]$, and $\tau_n = c_l(\log n)/n$ for a properly chosen absolute constant $c_l \geq 1$. The degree and interval for the approximation are $d = d_n = \Theta(\log n)$ and $[0, 1]$, respectively.

For our purpose, it suffices to approximate $f(x)$ to an order-$1/n$ error.

First-level truncation of $f(x)$ First, we show that only the lower-order part of $f(x)$ matters in the computation. By the definition of Bernstein polynomials and $|h_{m+1}(y)| \lesssim 1, \forall y \in [0,1]$,

$$\begin{aligned} B_m(h_{m+1}, \tau_n \cdot x) &= \mathop{\mathbb{E}}_{Y \sim \mathrm{bin}(m, \tau_n \cdot x)} \left[h_{m+1}\left(\frac{Y}{m}\right) \right] \\ &= \sum_{t=0}^{m} h_{m+1}\left(\frac{t}{m}\right) \cdot \Pr\left(\mathrm{bin}(m, \tau_n \cdot x) = t\right) \\ &= \left[\sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \Pr\left(\mathrm{bin}(m, \tau_n x) = t\right) \right] + O(\Pr(\mathrm{bin}(m, \tau_n x) > 4c_l \log^2 n)). \end{aligned}$$

Note that $m\tau_n \cdot x \le c_l \log^2 n$ for $x \in [0,1]$. Then, by standard binomial tail bounds, e.g, Lemma 6,

$$\Pr(\mathrm{bin}(m, \tau_n \cdot x) > 4c_l \log^2 n) \le e^{-c_l (\log^2 n)} \le \frac{1}{n^{\log n}} \le \frac{1}{n}.$$

Hence, we can *redefine* the function to approximate as

$$f(x) = \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \Pr\left(\mathrm{bin}(m, \tau_n x) = t\right) = \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} (\tau_n x)^t (1 - \tau_n x)^{m-t}.$$

A natural step to take is expending the polynomial function into its standard form.

$$\begin{aligned} f(x) &= \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} (\tau_n x)^t (1 - \tau_n x)^{m-t} \\ &= \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} (\tau_n x)^t \sum_{j=0}^{m-t} \binom{m-t}{j} (-\tau_n x)^{m-t-j} \\ &= \sum_{s=0}^{m} x^s \cdot \left(\tau_n^s \sum_{t=0}^{\min\{s, 4c_l \log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} \binom{m-t}{s-t} (-1)^{s-t} \right). \end{aligned}$$

For simplicity, let us denote the coefficient of x^s in $f(x)$ by C_s. Below, we bound the magnitude of

C_s for $s = 0, 1, \ldots, m$. Recall that $a \lesssim b$ represents $a = O(b)$ which hides only absolute constants, $|h_{m+1}(y)| \lesssim 1$ for all $y \in [0,1]$, and $\tau_n = c_l(\log n)/n$ for an absolute constant c_l. Then,

$$\begin{aligned}
|C_s| &= \left| \tau_n^s \sum_{t=0}^{\min\{s, 4c_l \log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t}\binom{m-t}{s-t}(-1)^{s-t} \right| \\
&\lesssim \left(\frac{c_l \log n}{n}\right)^s \sum_{t=0}^{s} \binom{m}{t}\binom{m-t}{s-t} \\
&\leq \left(\frac{c_l \log n}{n}\right)^s (2m)^s \\
&\leq \left(\frac{c_l \log n}{n}\right)^s (2n \log n)^s \\
&= \exp(\Theta(s \log\log n)).
\end{aligned}$$

Second-level truncation of $f(x)$ Following the above derivations, we can derive an alternative upper bound on C_s. This bound basically shows that for large s, the term corresponding to C_s is negligible. Specifically, consider any $s \geq 2(c_l e)^2 \log^4 n \geq (c_l e)^2 \log^4 n + 8c_l \log^2 n$ where

$c_l > 1$,

$$
\begin{aligned}
|C_s| &= \left| \tau_n^s \sum_{t=0}^{\min\{s,4c_l\log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t}\binom{m-t}{s-t}(-1)^{s-t} \right| \\
&\lesssim \left(\frac{c_l \log n}{n}\right)^s \sum_{t=0}^{4c_l\log^2 n} \binom{m}{t}\binom{m-t}{s-t} \\
&\le \left(\frac{c_l \log n}{n}\right)^s \sum_{t=0}^{4c_l\log^2 n} m^t \cdot \frac{m^{s-t}}{(s-t)!} \\
&\le \left(\frac{c_l \log n}{n}\right)^s (n\log n)^s \sum_{t=0}^{4c_l\log^2 n} \frac{1}{(s-t)!} \\
&\lesssim \frac{(c_l \log^2 n)^s}{(s-4c_l\log^2 n)!} \lesssim \frac{(c_l e \log^2 n)^s}{(s-4c_l\log^2 n)^{s-4c_l\log^2 n}} \\
&\le \frac{((c_l e)^2 \log^4 n)^{s/2}}{(s-4c_l\log^2 n)^{s-4c_l\log^2 n}} \\
&\le \frac{1}{((c_l e)^2\log^4 n)^{3\log^4 n}} \\
&\le \frac{1}{n^2 \log n} \le \frac{1}{mn}.
\end{aligned}
$$

Since $x \in [0,1]$, we can truncate $f(x)$ at degree $d_n^\star := 2(c_l e)^2 \log^4 n$ and *redefine* it as

$$
f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s,
$$

where C_s, as specified above, satisfies $|C_s| \lesssim \exp(\tilde{\Theta}(\log^4 n))$ and

$$
C_s = \tau_n^s \sum_{t=0}^{\min\{s,4c_l\log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t}\binom{m-t}{s-t}(-1)^{s-t}.
$$

This modification changes the value of $f(x)$ by at most $1/n$, for all $x \in [0,1]$.

Third-level truncation of $f(x)$ Now we evaluate each coefficient C_s to an error of $1/(nd_n^\star)$, so that we can compute $f(x)$ to an error of $1/n$, for all $x \in [0,1]$. This can be accom-

plished by computing every

$$C_{s,t} := h_{m+1}\left(\frac{t}{m}\right) \cdot \tau_n^s \binom{m}{t} \binom{m-t}{s-t} (-1)^{s-t}$$

to an $O(1/(nsd_n^\star))$ absolute error. Note that $C_{s,t}$ is a product of five terms, with each of them bounded by $m^s \leq \exp(\Theta(\log^5 n))$ in magnitude. Simple algebra further reduces our objective to approximating every term in the product to an $\exp(-\Theta(\log^5 n))$ error.

We analyze each term as follows: 1) computing $(-1)^{s-t}$ takes $O(\max\{\log s, \log t\}) = O(\log\log n)$ time; 2) computing the product of A integers of magnitude $\leq B$ takes $O((A\log B)^2)$ time, which can be achieved by recursively calculating the pairwise products [2]; 3) point 2) shows that we can compute $\binom{m}{t}$, $\binom{m-t}{s-t}$, and n^s *exactly* in ploylog(n) time; 4) now consider evaluating $(n\tau_n)^s = (c_l \log n)^s$: since $|a^s - b^s| \leq |a-b| \cdot s\max\{|a|,|b|\}^{s-1} \leq |a-b| \cdot O(\log^5 n)$ if $|a|, |b| \leq O(\log n)$, it suffices to compute $c_l \log n$ to an $\exp(-\Theta(\log^5 n))$ error, which can be performed in ploylog(n) time; 5) it remains to compute

$$h_{m+1}\left(\frac{t}{m}\right) = \log(m+1) - (t+1)\log(t+1) + t\log t,$$

to an $\exp(-\Theta(\log^5 n))$ error, which again takes ploylog(n) time.

Therefore, we can evaluate each $C_{s,t}$, and their sum C_s, to an error of $1/(nd_n^\star)$ in time ploylog(n). We can further define $C_s^\star$ as the closest integer multiple of $1/(nd_n^\star)$ [3] to C_s, and *redefine*

$$f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s^\star.$$

[2] We assume that computing the product two integers $\leq B$ takes $O(\log^2 B)$ time, achievable through the standard schoolbook "long multiplication". A more efficient integer-multiplication algorithm is the Harvey-Hoeven that takes only $\tilde{O}(\log B)$ time, yielding an $\tilde{O}(A\log B)$ complexity for the problem considered here.

[3] Assume that $d_n^\star$ is an integer. Otherwise, replace it by $\lceil d_n^\star \rceil$.

4.G.3 Lagrange Interpolation with Chebyshev Nodes

Recall that the degree of the min-max approximation polynomial is $d = d_n = \Theta(\log n)$. We initialize the Remez Algorithm by the Chebyshev nodes:

$$x_i := \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2i+1}{2(d+2)}\pi\right), i = 0, 1, \ldots, d_n + 1.$$

Then, for any integers $i \neq j \in [0, d+1]$,

$$\begin{aligned}
|x_i - x_j| &= \frac{1}{2}\left|\cos\left(\frac{2i+1}{2(d+2)}\pi\right) - \cos\left(\frac{2j+1}{2(d+2)}\pi\right)\right| \\
&= \left|\sin\left(\frac{i+j+1}{2(d+2)}\pi\right) \cdot \sin\left(\frac{i-j}{2(d+2)}\pi\right)\right| \\
&\geq \sin^2\left(\frac{\pi}{2(d+2)}\right) \geq \frac{1}{(d+2)^2}.
\end{aligned}$$

Now, consider the following function relating to the i-th Lagrange basis polynomial:

$$\ell_i(x) := \prod_{j \neq i}(x - x_j).$$

For any $\tau > 0$ and approximation sequence $\{x'_j\}_{j=0}^{d+1}$ in $[0,1]$ satisfying $|x_j - x'_j| \leq \tau$, denote by $\tilde{\ell}_i(x)$ the corresponding product $\prod_{j \neq i}(x - x'_j)$. Then, for any $x \in [0,1]$,

$$\begin{aligned}
|\ell_i(x) - \tilde{\ell}_i(x)| &\leq |\prod_{j \neq i}(x - x_j) - \prod_{j \neq i}(x - x'_j)| \\
&\leq \sum_{j \neq i} |(x - x_j) - (x - x'_j)| \prod_{j' < j, j' \neq i} |x - x_{j'}| \prod_{j' > j, j' \neq i} |x - x'_{j'}| \\
&\leq (d+1)\tau.
\end{aligned}$$

Under the same setting with $\tau < 1/(4(d+2)^2)$, the i-th Lagrange basis polynomial $L_i(x) := \ell_i(x)/\ell_i(x_i)$ and its approximation $\tilde{L}_i(x) := \tilde{\ell}_i(x)/\tilde{\ell}_i(x_i)$ differ by

$$\begin{aligned}
\left|L_i(x) - \tilde{L}_i(x)\right| &\le \left|\frac{\ell_i(x)}{\ell_i(x_i)} - \frac{\tilde{\ell}_i(x)}{\tilde{\ell}_i(x_i)}\right| \\
&= \left|\frac{\ell_i(x)\tilde{\ell}_i(x_i) - \tilde{\ell}_i(x)\ell_i(x_i)}{\ell_i(x_i)\tilde{\ell}_i(x_i)}\right| \\
&\le \left|(\tilde{\ell}_i(x_i) - \ell_i(x_i))\frac{\ell_i(x)}{\ell_i(x_i)\tilde{\ell}_i(x_i)}\right| + \left|(\ell_i(x) - \tilde{\ell}_i(x))\frac{\ell_i(x_i)}{\ell_i(x_i)\tilde{\ell}_i(x_i)}\right| \\
&\le \tau \cdot \exp(\tilde{\Theta}(\log n)).
\end{aligned}$$

Denote by $\mathcal{L}$ and $\tilde{\mathcal{L}}$ the Lagrange interpolation operator associated with $\{x_j\}_{j=0}^{d+1}$ and $\{x'_j\}_{j=0}^{d+1}$, respectively. Then for any $x \in [0,1]$, the interpolation polynomials of f differ by

$$\begin{aligned}
|\mathcal{L}[f](x) - \tilde{\mathcal{L}}[f](x)| &\le \sum_i |f(x_i)L_i(x) - f(x'_i)\tilde{L}_i(x)| \\
&\le \sum_i |(f(x_i) - f(x'_i))L_i(x) + f(x'_i)(L_i(x) - \tilde{L}_i(x))| \\
&\le \sum_i |L_i(x) \cdot \sum_{s=0}^{d_n^\star} (x_i^s - x_i'^s) \cdot C_s^\star| + \sum_i |f(x'_i)(L_i(x) - \tilde{L}_i(x))| \\
&\le \tau \cdot \exp(\tilde{\Theta}(\log^4 n)).
\end{aligned}$$

Set $\tau = \exp(-\tilde{\Theta}(\log^4 n))/n$ and recall that $E_d[g]$ denotes the best approximation error of the degree-d min-max polynomial over $[0,1]$. By the previous derivations and result of [HO20a],

for $T_d := 2 + \frac{2}{\pi}\log(d+1)$ and any $x \in [0,1]$,

$$\begin{aligned}
|\tilde{\mathcal{L}}[f](x) - B'_m(h,x)| &\le \frac{1}{n} + |\mathcal{L}[f](x) - B'_m(h,x)| \\
&\le \frac{1}{n} + |\mathcal{L}[f](x) - \mathcal{L}[B'_m(h,\cdot)](x) + \mathcal{L}[B'_m(h,\cdot)](x) - B'_m(h,x)| \\
&\le \frac{1}{n} + T_d \cdot (E_d[B'_m(h,\cdot)] + E_d[f] + E_d[B'_m(h,\cdot)]) + |f(x) - B'_m(h,x)| \\
&\le \frac{1}{n} + 3T_d \cdot E_d[B'_m(h,\cdot)] + (T_d+1) \max_{x\in[0,1]} |f(x) - B'_m(h,x)| \\
&\lesssim T_d \left(\frac{1}{n} + E_d[B'_m(h,\cdot)] \right) \\
&\lesssim \varepsilon \cdot \log\log n.
\end{aligned}$$

Therefore, if we compute each x_j to an $\exp(-\tilde{\Theta}(\log^4 n))$ error, the resulting polynomial $\tilde{\mathcal{L}}[f](x)$ approximates $B'_m(h,x)$ to an error of $O(\varepsilon \cdot \log\log n)$, for any $x \in [0,1]$. This yields a result only slightly weaker than that in Theorem 1, with the inequality being

$$L_{\hat{H}}(p,n) - L_{\hat{H}^E}(p, \varepsilon n \log n) \lesssim \varepsilon \cdot \log\log n \wedge \left(\frac{S_p}{n} + \frac{1}{n^{0.49}} \right).$$

Choose the approximation nodes $x'_j \in [0,1]$ to be integer multiples of $\exp(-\tilde{\Theta}(\log^4 n))$. Finally, we consider the time complexity of expanding $\tilde{\mathcal{L}}[f](x)$ into its standard form, which basically characterizes the time required for constructing the estimator. Note that

$$\tilde{\mathcal{L}}[f](x) = \sum_i f(x'_i) \cdot \frac{\prod_{j\neq i}(x - x'_j)}{\prod_{j\neq i}(x'_i - x'_j)}.$$

Since $x'_j \exp(\tilde{\Theta}(\log^4 n)) \in \mathbb{N}$ for any j and $f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s^\star$ with $C_s^\star$ being multiples of $1/(nd_n^\star)$, it takes $\mathrm{polylog}(n)$ time to evaluate $f(x'_i)$ and $\prod_{j\neq i}(x'_i - x'_j)$ exactly, with results expressed as rational numbers. In addition, computing each coefficient in the standard form of $\prod_{j\neq i}(x - x'_j)$ takes $O(2^d \cdot s^2) = \tilde{O}(\sqrt{n})$ [4] time. Hence, finding the explicit expression of the standard form of

[4] Recall that $d = c_s \log n$. Here we choose $c_s \le 1/2$.

$\tilde{\mathcal{L}}[f](x)$ takes $\tilde{O}(\sqrt{n}\log^2 n) = \tilde{O}(\sqrt{n})$ time. Let us denote this standard form by

$$\tilde{\mathcal{L}}[f](x) := \sum_{t=0}^{d+1} b_t \cdot x^t.$$

The small probability estimator is thus

$$\hat{\mathcal{V}}_S := \sum_{i\in[k]} \left(\sum_{t=1}^{d+2} \frac{b_{t-1}}{t} \cdot \frac{N_i^{\underline{t}}}{n^{\underline{t}}} \right) \cdot \mathbb{1}_{N_i' \le \frac{1}{\varepsilon}} \cdot \mathbb{1}_{N_i \lesssim \log n},$$

where N_i and N_i' are sample symbol counts in $[0,n]$. Note that computing each $N_i^{\underline{t}}$ or $n^{\underline{t}}$ takes $O(\log^2 n)$ time, and there are at most $O(\sqrt{n})$ distinct $(N_i, N_i' \lesssim 1/\varepsilon)$ pairs. Hence, we can evaluate the small-probability estimator in $\tilde{O}(n)$ time. In addition, the evaluation of the large-probability estimator is essentially the same as that of the empirical plug-in estimator. Consolidating these facts yields the desired near-linear-time computability.

4.G.4 Remez Algorithm with High Precision

Note that the first step of the Remez algorithm is initialization and will be executed only once. The last step of the algorithm serves as the initialization step for the next round of iteration. Exact evaluation of the initial nodes is not required in each round for convergence.

As shown by our previous discussion, it suffices to approximate the initial nodes to an accuracy of $\exp(-\text{polylog}(n))$, which takes $\text{polylog}(n)$ time for the first step. Denote by $x_0', \ldots x_{d+1}' \in [0,1]$ the initial nodes for a particular iteration and assume that $x_i'/\delta_n \in \mathbb{N}, i = 0, \ldots, d+1$.

We proceed to analyzing the second step of the Remez algorithm. According to Section 4.G.2, we will approximate the polynomial

$$f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s^\star,$$

where $d_n^\star = \Theta(\log^4 n)$ and $C_s^\star$'s are integer multiples of $1/(nd_n^\star)$ satisfying $|C_s^\star| \leq \exp(\tilde{\Theta}(\log^4 n))$. Computing the sequence of $f(x)$ values exactly for x_i''s takes $\text{polylog}(n)$ time. We can express each $f(x_i')$ as a rational number with both its nominator and denominator being at most $\exp(\text{polylog}(n))$. These claims clearly also hold for the evaluation of x^t at each x_i' with $t, j < d+2 = \Theta(\log n)$. Denote by $V_{b,E} := (b_0, \ldots, b_d, E)^T$ the vector of unknown variables. Multiplying both sides of each equation

$$b_0 + b_1 \cdot x_i' + \ldots + b_d \cdot x_i'^d + E \cdot (-1)^i = f(x_i')$$

by the least common multiple of the denominators of $x_i'^d$ and $f(x_i')$, we transform the second step to solving a system of linear equations in the form $AV_{b,E} = y$, where $A \in \mathbb{Z}_+^{(d+2)\times(d+2)}$ and $y \in \mathbb{Z}_+^{(d+2)\times 1}$ are matrices with entries bounded by $\exp(\text{polylog}(n))$. If the initial nodes x_j''s are distinct and sorted accordingly, the system $AV_{b,E} = y$ has a unique solution. Utilizing the algorithm proposed by [Dix82], we can solve this system in time $\tilde{O}((d+2)^3 \log(\|A\| + \|y\|)) = \text{polylog}(n)$ where $\|\cdot\|$ represents the maximum entry in absolute value.

Once we obtain the coefficient vector $V_{b,E}$, Step 3 of the algorithm takes $\text{polylog}(n)$ time to form the approximation polynomial

$$P(x) := b_0 + b_1 \cdot x + \ldots + b_d \cdot x^d.$$

The fourth step of the Remez algorithm calls for computing the local extrema of the error function

$$\mathcal{E}(x) := P(x) - f(x)$$

over the $d+2$ sign-invariant regions. Noting that $\mathcal{E}(x)$ is a degree $d_n^\star$ polynomial, it suffices to approximate all the real roots of its derivative $\mathcal{E}'(x)$ to an $\exp(-\text{polylog}(n))$ accuracy.

To do this, we first transform $\mathcal{E}'(x)$ to a polynomial with integer coefficients of size $\exp(\text{polylog}(n))$. Then, we apply the quadratic interval refinement algorithm [Abb14] to ap-

proximate the real roots of the transformed polynomial. Shown in the paper of [Ker09], for a degree-d square-free polynomial with integer coefficients bounded by 2^σ in absolute value, an ε-accuracy approximation of the real roots using this algorithm requires a time complexity of $\tilde{O}(d^4\sigma^2 + d^3\log(1/\varepsilon))$. For the task considered here, this again converts to a time complexity of $\mathrm{polylog}(n)$.

Finally, we can view Step 5 as the initialization step in the next iteration, implying a per-iteration complexity of $\mathrm{polylog}(n)$ for the Remez algorithm. Note that quantity E corresponds to a lower bound on the max approximation error of each iteration. As for the number of iterations, [Vei60] essentially shows that under differentiability, this process has a quadratic convergence. More specifically, let E_ν denote the error bound E of the the ν-th iteration, then $\{E_\nu\}_{\nu\geq 1}$ converges to the optimal degree-d approximation error $E_d[f]$ with

$$|E_d[f] - E_\nu| \lesssim (E_d[f] - E_{\nu-1})^2.$$

It takes only $\mathrm{polylog}(n)$ iterations for E to converge to the $\exp(-\mathrm{polylog}(n))$-neighborhood of its limit $E_d[f]$. Therefore, the total time required for computing the approximation polynomial with Remez algorithm is $O(\mathrm{polylog}(n))$. Consolidating this with the reasoning in the last section shows that our estimator can be evaluated in time near-linear in n. On the practical side, see [PT09, Tre13] for an optimized Matlab implementation of the Remez algorithm.

4.H A Refined Estimator for Shannon Entropy

In this section, we replacing the function $h_n(x)$ employed in Section 4.C by a much finer approximation of $B_n(h,x)$. Through this refinement, we establish the full version of Theorem 1.

To begin with, we define the following two *f-functions* for $z \in [0,\infty]$:

$$f_1(z) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(z)}[h(X)] = -e^{-z}\sum_{j=1}^{\infty}\frac{z^j}{j!}j\log j$$

and

$$f_2(z) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(z)}[h(X+1)] = -e^{-z}\sum_{j=1}^{\infty}\frac{z^j}{j!}(j+1)\log(j+1).$$

4.H.1 Relating f-functions to Bernstein Approximation Errors

For $x \in [0,1]$, set $z = z(x) := nx$. The following lemma relates $f_1(z)$ and $f_2(z)$ to the Bernstein approximation error of h_{n+1}, that is, $h_{n+1}(x) - B_n(h_{n+1},x)$.

Lemma 13. *For any $x \in [0,\log^4 n/n]$,*

$$h_{n+1}(x) - B_n(h_{n+1},x) = (h(z+1) - f_2(z)) - (h(z) - f_1(z)) + \tilde{O}\left(\frac{1}{n}\right).$$

As a corollary, for any sufficiently large n and $x \in I_n = [0,\tau_n := c_l(\log n)/n]$,

$$h_{na}(x) - B_{na-1}(h_{na},x) = (h(z+1) - f_2(z)) - (h(z) - f_1(z)) + \tilde{O}\left(\frac{1}{na-1}\right).$$

Since $1/(na-1) \le \min\{1/\log n, S_p/n\}$, the last term on the right-hand side is negligible. These results, together with the function-wise triangle inequality on w_φ^2, further reduce the desired inequality

$$w_\varphi^2(B_{na-1}(h_{na},\tau_n\cdot x), d_n^{-1}) \lesssim \varepsilon$$

to bounds in the form of

$$w_\varphi^2(g(x), d_n^{-1}) \lesssim \varepsilon,$$

for function $g(x)$ being $h_{na}(\tau_n\cdot x)$, $h(z(x))$, $h(z(x)+1)$, $f_1(z(x))$, and $f_2(z(x))$, respectively.

Proof. Let $h_{-1}(x) := h(x+n^{-1})$. By the linearity of expectation,

$$\begin{aligned} h_{n+1}(x) - B_n(h_{n+1},x) &= n\left(h_{-1}(x) - h(x) - B_n(h_{-1},x) + B_n(h,x)\right) \\ &= n\left(h_{-1}(x) - B_n(h_{-1},x)\right) - n\left(h(x) - B_n(h,x)\right). \end{aligned}$$

Note that $z = nx$ implies $z \in [0, \log^4 n]$. Hence, we have

$$\begin{aligned} n\left(h_{-1}(x) - B_n(h_1,x)\right) &= -(nx+1)\log\left(\frac{nx+1}{n}\right) + \sum_{j=0}^{n}(j+1)\log\left(\frac{j+1}{n}\right)\binom{n}{j}x^j(1-x)^{n-j} \\ &= -(z+1)\log\left(\frac{z+1}{n}\right) + \sum_{j=0}^{n}(j+1)\log\left(\frac{j+1}{n}\right)\binom{n}{j}z^j\frac{(n-z)^{n-j}}{n^n} \\ &= -(z+1)\log(z+1) + \left(1-\frac{z}{n}\right)^n\sum_{j=0}^{n}(j+1)\log(j+1)\binom{n}{j}z^j(n-z)^{-j} \\ &= -(z+1)\log(z+1) + \left(1-\frac{z}{n}\right)^n\sum_{j=0}^{n}(j+1)\log(j+1)\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\ &= -(z+1)\log(z+1) + e^{-z}\sum_{j=0}^{\infty}\frac{z^j}{j!}(j+1)\log(j+1) + \tilde{O}\left(\frac{1}{n}\right) \\ &= h(z+1) - f_2(z) + \tilde{O}\left(\frac{1}{n}\right). \end{aligned}$$

The second last equality is the most non-trivial step. In order to establish this equality, we will need the following three inequalities (assume $z \in [0, \log^4 n]$ and $n \gg 1$).

Inequality 1:

$$\begin{aligned}
0 &\le \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} (j+1)\log(j+1)\,\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\
&= \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} (j+1)\log(j+1)\,\frac{n^{\underline{j}}}{2^j(n-z)^j}\frac{(2z)^j}{j!} \\
&\le e^{-z} \sum_{j=\log^5 n+1}^{n} (j+1)\log(j+1)\,\frac{(2z)^j}{j!} \\
&\le e^{-z} \sum_{j=\log^5 n+1}^{n} 2j(j-1)\frac{(2z)^j}{j!} \\
&\le 8z^2 e^{-z} \sum_{j=\log^5 n-1}^{n} \frac{(2z)^j}{j!} \\
&\le 8(\log^8 n)\Pr(\mathrm{Poi}(2z)\ge \log^5 n-1) \\
&\le \frac{1}{n}.
\end{aligned}$$

Inequality 2:

$$0 \le e^{-z} \sum_{j=\log^5 n+1}^{\infty} \frac{z^j}{j!}(j+1)\log(j+1) = 2(\log^8 n)\Pr(\mathrm{Poi}(2z)\ge \log^5 n-1) \le \frac{1}{n}.$$

Inequality 3: For any $j \le \log^5 n$,

$$\begin{aligned}
\left| e^{-z} - \left(1-\frac{z}{n}\right)^n \frac{n^{\underline{j}}}{n^j} \left(1-\frac{z}{n}\right)^{-j} \right| &= \left| e^{-z} - \left(1-\frac{z}{n}\right)^n \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le \left| e^{-z} - \left(1-\frac{z}{n}\right)^n \right| + \left(1-\frac{z}{n}\right)^n \left| 1 - \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le e^{-z}\frac{z^2}{n} + e^{-z} \left| 1 - \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le e^{-z}\frac{z^2}{n} + e^{-z} \left(\left| 1 - \frac{n^j}{(n-z)^j} \right| \vee \left| 1 - \frac{(n-\log^5 n)^j}{(n-z)^j} \right| \right) \\
&\le e^{-z}\frac{z^2}{n} + e^{-z} \left(\left| \frac{zj}{n-z(j+1)} \right| \vee \left| \frac{(\log^5 n)\, j}{n-z} \right| \right) \\
&\le e^{-z}\frac{2\log^{10} n}{n}.
\end{aligned}$$

Note that Inequality 3 further implies

$$\begin{aligned}
&\left| e^{-z} \sum_{j=0}^{\log^5 n} \frac{z^j}{j!}(j+1)\log(j+1) - \left(1-\frac{z}{n}\right)^n \sum_{j=0}^{\log^5 n} (j+1)\log(j+1)\, \frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!} \left(1-\frac{z}{n}\right)^{-j} \right| \\
&\le \frac{2\log^{10} n}{n} \cdot e^{-z} \sum_{j=0}^{\log^5 n} \frac{z^j}{j!}(2j(j-1)) \\
&\le \frac{2\log^{10} n}{n} \cdot 2z^2 \\
&\le \frac{4\log^{18} n}{n}.
\end{aligned}$$

This, together with Inequality 1 and 2, proves the desired equality. The same reasoning also gives

$$n\,(h(x) - B_n(h,x)) = -z\log z + e^{-z} \sum_{j=1}^{\infty} \frac{z^j}{j!} j\log j + \tilde{O}\left(\frac{1}{n}\right),$$

which completes the proof. □

For any $x \in I_n$, let $z_1 = (na-1)x$, then $z_1 \in I_n' := [0, ac_l \log n]$. Therefore, by Lemma 13,

$$h_{na}(x) - B_{na-1}(h_{na}, x) = (h(z_1+1) - f_2(z_1)) - (h(z_1) - f_1(z_1)) + \tilde{O}\left(\frac{1}{n}\right).$$

In the next section, we approximate function $f_1(z)$ over I_n' with a degree-d polynomial.

4.H.2 Approximating $f_1(z)$

Consider the first function

$$f_1(z) = -e^{-z} \sum_{j=1}^{\infty} \frac{z^j}{j!} j \log j.$$

We want to approximate f_1 with a low-degree polynomial and bound the corresponding error. For this purpose, we establish some basic properties of $f_1(z)$ as follows.

Properties of $f_1(z)$

Property 1: The function $f_1(z)$ is a continuous function over $[0, \infty)$, and $f_1(0) = 0$.

Property 2: For all $z \geq 0$, the value of $f_1(z)$ is non-negative.

Property 3: Denote $u(y) := (y+2)\log(y+2) + y\log y - 2(y+1)\log(y+1)$. Then, for any $z \geq 0$,

$$f_1{}''(z) = -e^{-z} \sum_{t=0}^{\infty} \frac{z^t}{t!} \cdot u(t) \text{ and } -\log 4 \leq f_1{}''(z) < 0.$$

Proof. We begin by establishing the equality.

$$\begin{aligned}
-f_1{}''(z) &= e^{-z} \sum_{t=1}^{\infty} \frac{(t-1)t^2 z^{t-2} \log(t)}{t!} - 2e^{-z} \sum_{t=1}^{\infty} \frac{t^2 z^{t-1} \log(t)}{t!} + e^{-z} \sum_{t=1}^{\infty} \frac{t z^t \log(t)}{t!} \\
&= e^{-z} \sum_{t=0}^{\infty} \frac{z^t (t+2)\log(t+2)}{t!} - 2e^{-z} \sum_{t=0}^{\infty} \frac{z^t (t+1)\log(t+1)}{t!} + e^{-z} \sum_{t=0}^{\infty} \frac{t z^t \log(t)}{t!} \\
&= e^{-z} \sum_{t=0}^{\infty} \frac{z^t}{t!} \cdot u(t).
\end{aligned}$$

To prove the inequality, we utilize the following lemma.

Lemma 14. *For any $t \geq 0$,*

$$\frac{\log 4}{t+1} \geq u(t) \geq \frac{1}{t+1}.$$

By Lemma 14, we obtain

$$0 < e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot\frac{1}{t+1} \leq e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot u(t) = -{f_1}''(z) \leq e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot\frac{\log 4}{t+1} = (\log 4)\frac{1-e^{-z}}{z} \leq \log 4.$$

The proof of the lemma follows by standard algebraic calculations and is omitted. □

Property 4: For $z > 0$,

$$0 \leq \frac{{f_1}''(z)}{h''(z)} \leq \log 4.$$

Proof. Recall that $h(z) = -z\log z$. Therefore, $h''(z) = -1/z$ and

$$\begin{aligned}
0 &\leq \frac{{f_1}''(z)}{h''(z)} \\
&= e^{-z}\sum_{t=0}^{\infty}\frac{z^{t+1}}{t!}\cdot u(t) \\
&\leq e^{-z}\sum_{t=0}^{\infty}\frac{z^{t+1}}{t!}\cdot\frac{\log 4}{t+1} \\
&\leq (\log 4)(1-e^{-z}) \\
&\leq \log 4,
\end{aligned}$$

where the third step follows by Lemma 14. □

Moduli of Smoothness

In this section, we introduce some notable results in approximation theory [DT12] that are crucial for our simplification of the problem. Let $\varphi(x) := \sqrt{x(1-x)}$. For any function

$f : [0,1] \to \mathbb{R}$, *the first- and second- order Ditzian-Totik moduli of smoothness* quantities of f are

$$w^1_\varphi(f,t) := \sup\left\{ |f(u) - f(v)| : 0 \le u, v \le 1, |u-v| \le t \cdot \varphi\left(\frac{u+v}{2}\right) \right\},$$

and

$$w^2_\varphi(f,t) := \sup\left\{ \left| f(u) + f(v) - 2f\left(\frac{u+v}{2}\right) \right| : 0 \le u, v \le 1, |u-v| \le 2t \cdot \varphi\left(\frac{u+v}{2}\right) \right\},$$

respectively. Let P_d denote the collection of polynomials with real coefficients and degree at most d. For any $d \in \mathbb{Z}^+$, interval $I \subset \mathbb{R}$, and function $f : I \to \mathbb{R}$, denote by

$$E_d[f, I] := \min_{\tilde{f} \in P_d} \max_{x \in I} |f(x) - \tilde{f}(x)|$$

the *best approximation error* of the degree-d min-max polynomial of f over I. For a bounded domain I, we can always shift and rescale f to make it a real function over $[0,1]$. Hence, without loss of generality, it suffices to consider and analyze $E_d[f] := E_d[f, [0,1]]$.

The connection between the best polynomial-approximation error $E_d[f]$ of a continuous function f and the second order Ditzian-Totik moduli of smoothness $w^2_\varphi(f,t)$ is established in the following lemma [DT12].

Lemma 15. *There are absolute constants C_1 and C_2 such that for any continuous function f over $[0,1]$ and $d > 2$,*

$$E_d[f] \le C_1 w^2_\varphi(f, d^{-1}),$$

and

$$\frac{1}{d^2} \sum_{t=0}^{d} (t+1) E_t[f] \ge C_2 w^2_\varphi(f, d^{-1}).$$

The above lemma shows that the second order smoothness quantity $w^2_\varphi(f,\cdot)$ essentially

characterizes $E_\cdot[f]$, and thus transforms the problem of showing

$$|\tilde{h}_m(x) - B_{m-1}(h_m, x)| \lesssim \varepsilon, \forall x \in I_n,$$

to that of establishing

$$w_\varphi^2(B_{m-1}(h_m, \tau_n \cdot x), d_n^{-1}) \lesssim \varepsilon,$$

where $\tau_n = \Theta(\log n/n)$ and $d_n = \Theta(\log n)$ by definition.

Bounding Errors in Approximating $f_1(x)$

For simplicity, define $x' := (ac_l \log n) \cdot x$ and consider the function

$$f_{1'}(x) := f_1((ac_l \log n) \cdot x).$$

Under proper scaling, approximating $f_1(x')$ over $I_n' = [0, ac_l \log n]$ is equivalent to approximating $f_{1'}(x)$ over $[0,1]$. By Lemma 15, it suffices to bound $w_\varphi^2(f_{1'}, \cdot)$ for our purpose.

In particular, we know that

$$\min_{g \in P_d} \max_{x \in I_n'} |f_1(x) - g(x)| = E_d[f_{1'}] \leq C_1 w_\varphi^2(f_{1'}, d^{-1}).$$

By definition, $w_\varphi^2(f_{1'}, d^{-1})$ is the solution to the following optimization problem.

$$\sup_{u,v} \left| f_{1'}(u) + f_{1'}(v) - 2f_{1'}\left(\frac{u+v}{2}\right) \right|$$

subject to

$$0 \leq u, v \leq 1, |u - v| \leq \frac{2}{d} \cdot \varphi\left(\frac{u+v}{2}\right).$$

First, consider the optimization constraints. Analogous to the arguments in [JVHW15], we define

$M := (u+v)/2$ and $\delta := d^{-1}\sqrt{1/M-1}$. The feasible region can be expressed as

$$[M-d^{-1}\sqrt{M(1-M)}, M+d^{-1}\sqrt{M(1-M)}] \cap [0,1] = [M-\delta M, M+\delta M] \cap [0,1].$$

By Property 3 in Section 4.H.2, $f_1(x')$, or equivalently $f_{1'}(x)$, is a strictly concave function. Therefore, the maximum of $|f(u)+f(v)-2f(u+v/2)|$ is attained at the boundary of the feasible region.

Note that

$$M-d^{-1}\sqrt{M(1-M)} \geq 0 \iff M \geq \frac{1}{d^2+1}$$

and

$$M+d^{-1}\sqrt{M(1-M)} \leq 1 \iff M \leq \frac{d^2}{d^2+1}.$$

We need to consider only three cases:

Case 1:

$$u = 0, v = 2M, M \in [0, 1/(d^2+1)].$$

Case 2:

$$u = 2M-1, v = 1, M \in [d^2/(d^2+1), 1].$$

Case 3:

$$u = M-\delta M, v = M+\delta M, M \in [1/(d^2+1), d^2/(d^2+1)].$$

To facilitate the discussions, we utilize the following lemma.

Lemma 16. *Let $f \in C^1([a,b])$ have second order derivative in (a,b). There exists $c \in (a,b)$ such*

that

$$f(a)+f(b)-2f\left(\frac{a+b}{2}\right)=\frac{1}{4}(b-a)^2\cdot f''(c).$$

We begin with Case 1. By the Lemma 16, there exists $c\in(0,2/(d^2+1))$ satisfying

$$\left|f_{1'}(0)+f_{1'}\left(\frac{2}{d^2+1}\right)-2f_{1'}\left(\frac{1}{d^2+1}\right)\right|\leq\frac{1}{4}\cdot\left(\frac{2}{d^2+1}\right)^2\left|f_{1'}{}''(c)\right|=\left(\frac{1}{d^2+1}\right)^2\left|f_{1'}{}''(c)\right|.$$

By the definition of function $f_{1'}$,

$$|f_{1'}{}''(x)|=|(ac_l\log n)^2 g_1''((ac_l\log n)\cdot x)|\leq(\log 4)(ac_l\log n)^2.$$

Therefore, we obtain

$$\left(\frac{1}{d^2+1}\right)^2\left|f_{1'}{}''(c)\right|\lesssim\varepsilon^2.$$

This, together with an analogous argument on Case 2, implies that the objective value is bounded by $O(\varepsilon^2)$ in both cases. It remains to analyze Case 3. We proceed by considering two regimes:

Regime 1: If $M\leq 4/(d^2+1)$, then $|u-v|=2d^{-1}\sqrt{M(1-M)}\leq 4/d^2$. The above reasoning again shows that

$$\left|f_{1'}(u)+f_{1'}(v)-2f_{1'}\left(\frac{u+v}{2}\right)\right|\lesssim\varepsilon^2.$$

Regime 2: If $4/(d^2+1)\leq M\leq d^2/(d^2+1)$,

$$M-\delta M=M\left(1-\frac{\sqrt{M^{-1}-1}}{d}\right)\geq M\left(1-\frac{\sqrt{(d^2+1)-4}}{2d}\right)\geq\frac{M}{2}.$$

By Lemma 16, there exists $c \in (M-\delta M, M+\delta M) \subseteq (M/2, 3M/2)$ satisfying

$$\left| f_{1'}(u) + f_{1'}(v) - 2f_{1'}\left(\frac{u+v}{2}\right) \right| \leq \frac{1}{4} \cdot \left(2\frac{1}{d}\sqrt{M(1-M)}\right)^2 \left|f_{1'}{}''(c)\right|.$$

Then, by Property 4 in Section 4.H.2,

$$|f_{1'}{}''(c)| = |(ac_l \log n)^2 f_1''((ac_l \log n)\cdot c)| \leq (ac_l \log n)^2 (\log 4) \cdot \frac{1}{(ac_l \log n)\cdot c} \leq (\log 8) \cdot \frac{ac_l \log n}{M}.$$

This bound immediately implies

$$\frac{1}{4} \cdot \left(\frac{2}{d}\sqrt{M(1-M)}\right)^2 \cdot \left|f_{1'}{}''(c)\right| \leq \frac{1}{d^2} M(1-M) \cdot (\log 8) \cdot \frac{ac_l \log n}{M} \leq (\log 8) \cdot \frac{c_l \varepsilon}{c_s^2}.$$

Consolidating the previous results yields

$$\min_{g \in P_d} \max_{x \in I_n'} |f_1(x) - g(x)| \lesssim \varepsilon.$$

For function f_2, an analogous argument shows that

$$\min_{g \in P_d} \max_{x \in I_n'} |f_2(x) - g(x)| \lesssim \varepsilon.$$

In the next section, we apply these inequalities to analyze our refined entropy estimator.

4.H.3 Proving Theorem 1: A Refined Entropy Estimator

We aim to approximate $B_{na-1}(h_{na}, x) - h_{na}(x)$ over $I_n = [0, c_l \log n/n]$ by a degree-d polynomial. By Lemma 13, for any $x \in I_n$ and $z_1 := (na-1)x \in I_n' = [0, ac_l \log n]$,

$$h_{na}(x) - B_{na-1}(h_{na}, x) = (h(z_1+1) - f_2(z_1)) - (h(z_1) - f_1(z_1)) + \tilde{O}\left(\frac{1}{n}\right).$$

By the results in [Kor91],

$$\min_{g\in P_d}\max_{x\in I'_n}|h(x)-g(x)| = (ac_l\log n)\min_{g\in P_d}\max_{x\in[0,1]}|h(x)-g(x)| \lesssim \frac{ac_l\log n}{(c_s\log n)^2}\lesssim \varepsilon$$

and

$$\min_{g\in P_d}\max_{x\in I'_n}|h(x+1)-g(x)|\lesssim \varepsilon.$$

Combining these bounds with the last two inequalities in the previous section, we obtain

$$\min_{g\in P_{d-1}}\max_{x\in I_n}|(h_{na}(x)-B_{na-1}(h_{na},x))-g(x)|\lesssim \varepsilon.$$

Denote by $\tilde{g}(x)$ the min-max polynomial that achieves this minimal error. By the derivations in Section 4.C.2, the degree-$(d-1)$ polynomial $\tilde{h}_{na}(x)$ satisfies

$$\max_{x\in I_n}|h_{na}(x)-\tilde{h}_{na}(x)|\lesssim \varepsilon.$$

Denote $\tilde{h}^*(x) := -\tilde{g}(x)+\tilde{h}_{na}(x)$, and note that by definition, $B'_{na}(h,x) = B_{na-1}(h_{na},x)$. Then, the triangle inequality implies

$$\max_{x\in I_n}|B'_{na}(h,x)-\tilde{h}^*(x)| = \max_{x\in I_n}|B_{na-1}(h_{na},x)-\tilde{h}^*(x)|\lesssim \varepsilon.$$

By the triangle inequality of integrals, the degree-d polynomial

$$\tilde{H}^*(x) := \int_0^x \tilde{h}^*(t)dt$$

approximating $B_{na}(h,x)$ possesses the following pointwise error guarantee.

Lemma 17. *For any $x\in I_n$,*

$$|B_{na}(h,x)-\tilde{H}^*(x)|\lesssim x\varepsilon.$$

Hence, $\tilde{H}^*(x)$ is a degree-d polynomial that well approximates $B_{na}(h,x)$ pointwisely.

Next, we argue that the coefficients of $\tilde{H}^*(x)$ can not be too large. For notational convenience, write $\tilde{h}^*(x) := \sum_{v=0}^{d-1} a_v x^v$. By Corollary 2, for any $x \in I_n$,

$$|h_{na}(x) - B_{na-1}(h_{na},x)| \leq 1.$$

Furthermore, $h_{na}(x)$ is an increasing function over I_n, and thus

$$|h_{na}(x)| = \max\left\{|h_{na}(0)|, h_{na}\left(\frac{c_l(\log n)}{n}\right)\right\} \lesssim \log n.$$

Therefore, for any $x \in I_n$,

$$|\tilde{h}^*(x)| \lesssim \log n.$$

The boundedness of $\tilde{h}^*(x)$ implies that its coefficients cannot be too large:

$$|a_v| \lesssim \left(2^{4.5d}\log n\right)\left(\frac{n}{c_l \log n}\right)^v.$$

Write $\tilde{H}^*(x)$ as $\tilde{H}^*(x) = \sum_{t=1}^{d} a_t' x^t$. Then, by $\tilde{H}^*(x) = \int_0^x \tilde{h}^*(t)dt$ and the bound on $|a_v|$,

$$|a_t'| \lesssim 2^{4.5d}\left(\frac{n}{c_l \log n}\right)^{t-1}.$$

The construction of the new entropy estimator follows by replacing $\tilde{H}_{na}(x)$ by $\tilde{H}^*(x)$ in Section 4.D. The rest of the proof is also similar to that in the main paper and thus omitted.

4.I Competitive Estimators for General Additive Properties

Consider an arbitrary real function $f : [0,1] \to \mathbb{R}$. Without loss of generality, we will assume that $f(0) = 0$. According to the derivations in Section 4.C, we can write $B'_n(f,x)$ as

$$B'_n(f,x) := \sum_{j=0}^{n-1} n\left(f\left(\frac{j+1}{n}\right) - f\left(\frac{j}{n}\right)\right)\binom{n-1}{j}x^j(1-x)^{(n-1)-j}.$$

Our aim to approximate $B'_{na}(f,x)$ with a low degree polynomial. For simplicity, we assume that f is a 1-Lipschitz function. For $x \in [0,1]$, set $z = nx$, and define $g_{n+1}(j) := (n+1)f\left(\frac{j}{n+1}\right)$,

$$f_{1,n+1}(z) := e^{-z}\sum_{j=0}^{\infty} g_{n+1}(j+1)\frac{z^j}{j!},$$

and

$$f_{2,n+1}(z) := e^{-z}\sum_{j=0}^{\infty} g_{n+1}(j)\frac{z^j}{j!}.$$

The following lemma relates $f_{1,n+1}(z)$ and $f_{2,n+1}(z)$ to $B'_{n+1}(f,x)$.

Lemma 18. *For any $x \in [0, \log^4 n/n]$ and $z = nx$,*

$$B'_{n+1}(f,x) = f_{1,n+1}(z) - f_{2,n+1}(z) + \tilde{O}\left(\frac{1}{n}\right).$$

Proof. Note that $z = nx$ implies $z \in [0, \log^4 n]$. Hence, we have

$$\begin{aligned}
\sum_{j=0}^{n}(n+1)f\left(\frac{j+1}{n+1}\right)\binom{n}{j}x^j(1-x)^{n-j} &= \sum_{j=0}^{n} g_{n+1}(j+1)\binom{n}{j}z^j\frac{(n-z)^{n-j}}{n^n} \\
&= \left(1-\frac{z}{n}\right)^n \sum_{j=0}^{n} g_{n+1}(j+1)\binom{n}{j}z^j(n-z)^{-j} \\
&= \left(1-\frac{z}{n}\right)^n \sum_{j=0}^{n} g_{n+1}(j+1)\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\
&= e^{-z}\sum_{j=0}^{\infty} g_{n+1}(j+1)\frac{z^j}{j!} + \tilde{O}\left(\frac{1}{n}\right) \\
&= f_{1,n+1}(z) + \tilde{O}\left(\frac{1}{n}\right).
\end{aligned}$$

The second last equality is the most non-trivial step. In order to establish this equality, we will need the following three inequalities (assume $z \in [0, \log^4 n]$ and $n \gg 1$).

Inequality 1:

$$\begin{aligned}
0 &\le \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} |g_{n+1}(j+1)|\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\
&= \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} (j+1)\frac{n^{\underline{j}}}{2^j(n-z)^j}\frac{(2z)^j}{j!} \\
&\le e^{-z}\sum_{j=\log^5 n+1}^{n} (j+1)\frac{(2z)^j}{j!} \\
&\le e^{-z}\sum_{j=\log^5 n+1}^{n} 2j(j-1)\frac{(2z)^j}{j!} \\
&\le 8z^2e^{-z}\sum_{j=\log^5 n-1}^{n}\frac{(2z)^j}{j!} \\
&\le 8(\log^8 n)\Pr(\mathrm{Poi}(2z) \ge \log^5 n - 1) \\
&\le \frac{1}{n}.
\end{aligned}$$

Inequality 2:

$$0 \le e^{-z} \sum_{j=\log^5 n+1}^{\infty} |g_{n+1}(j+1)| \frac{z^j}{j!} \le e^{-z} \sum_{j=\log^5 n+1}^{\infty} (j+1) \frac{z^j}{j!} \le \frac{1}{n}.$$

Inequality 3: For any $j \le \log^5 n$,

$$\begin{aligned}
\left| e^{-z} - \left(1-\frac{z}{n}\right)^n \frac{n^{\underline{j}}}{n^j} \left(1-\frac{z}{n}\right)^{-j} \right| &= \left| e^{-z} - \left(1-\frac{z}{n}\right)^n \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le \left| e^{-z} - \left(1-\frac{z}{n}\right)^n \right| + \left(1-\frac{z}{n}\right)^n \left| 1 - \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left| 1 - \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left(\left| 1 - \frac{n^j}{(n-z)^j} \right| \vee \left| 1 - \frac{(n-\log^5 n)^j}{(n-z)^j} \right| \right) \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left(\left| \exp\left(\frac{zj}{n-z}\right) - 1 \right| \vee \left| \frac{(\log^5 n - z) j}{n-z} \right| \right) \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left(\left| \frac{zj}{n-z(j+1)} \right| \vee \left| \frac{(\log^5 n) j}{n-z} \right| \right) \\
&\le e^{-z} \frac{2\log^{10} n}{n}.
\end{aligned}$$

Note that Inequality 3 further implies

$$\begin{aligned}
&\left| e^{-z} \sum_{j=0}^{\log^5 n} \frac{z^j}{j!} g_{n+1}(j+1) - \left(1-\frac{z}{n}\right)^n \sum_{j=0}^{\log^5 n} g_{n+1}(j+1) \frac{n^{\underline{j}}}{n^j} \frac{z^j}{j!} \left(1-\frac{z}{n}\right)^{-j} \right| \\
&\le \frac{2\log^{10} n}{n} \cdot e^{-z} \sum_{j=0}^{\log^5 n} \frac{z^j}{j!} (j+1) \\
&\le \frac{2\log^{10} n}{n} \cdot (1+2z) \\
&\le \frac{5\log^{14} n}{n}.
\end{aligned}$$

This, together with Inequality 1 and 2, proves the desired equality. The same reasoning also gives

$$\sum_{j=0}^{n}(n+1)f\left(\frac{j}{n+1}\right)\binom{n}{j}x^j(1-x)^{n-j} = f_{2,n+1}(z) + \tilde{O}\left(\frac{1}{n}\right),$$

which completes the proof. □

By slightly abusing the notation, we redefine $z := (na-1)x$. Lemma 18 immediately implies that for any $x \in I_n = [0, c_l(\log n)/n] \subseteq [0, (\log^4(na-1))/(na-1)]$,

$$B'_{na}(f,x) = f_{1,na}(z) - f_{2,na}(z) + \tilde{O}\left(\frac{1}{na}\right).$$

Note that $z \in I'_n = [0, ac_l \log n]$ in this case. Define $t_{na}(z) := f_{1,na}(z) - f_{2,na}(z)$ and $r_{na}(j) := g_{na}(j+2) + g_{na}(j) - 2g_{na}(j+1)$. Then, direct calculation yields

$$\begin{aligned}
t''_{na}(z) &= e^{-z}\sum_{j=0}^{\infty} r_{na}(j+1)\frac{z^j}{j!} - e^{-z}\sum_{j=0}^{\infty} r_{na}(j)\frac{z^j}{j!} \\
&= e^{-z}\sum_{j=0}^{\infty} r_{na}(j+1)\frac{z^j}{j!} - e^{-z}r_{na}(0) - \sum_{j=0}^{\infty} r_{na}(j+1)\frac{z^{j+1}}{(j+1)!} \\
&= e^{-z}\sum_{j=0}^{\infty} r_{na}(j+1)\left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!}\right) - e^{-z}r_{na}(0).
\end{aligned}$$

Since f is 1-Lipschitz, we obtain $|r_{na}(j)| \le 2$. Therefore, for any $z \in I'_n$,

$$|t''_{na}(z)| \le e^{-z}\sum_{j=0}^{\infty} |r_{na}(j+1)|\left(\frac{z^j}{j!} + \frac{z^{j+1}}{(j+1)!}\right) + e^{-z}|r_{na}(0)| \le 6.$$

We can bound each summand in the expression of t''_{na} by the following lemma.

Lemma 19. *For any $j \ge 1$ and $z \ge 0$, we have*

$$\left|e^{-z}\left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!}\right)\right| \le \frac{1}{\sqrt{2\pi}((j+1) - \sqrt{j+1})}$$

and

$$\left| e^{-z} \left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!} \right) \right| \leq \frac{5}{z}.$$

Proof. For the ease of exposition, denote

$$q_1(z) := e^{-z} \left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!} \right).$$

Then, the derivative of $q_1(z)$ is

$$\begin{aligned} q_1'(z) &= -e^{-z}\frac{z^j}{j!} + e^{-z}\frac{z^{j-1}}{(j-1)!} + e^{-z}\frac{z^{j+1}}{(j+1)!} - e^{-z}\frac{z^j}{j!} \\ &= e^{-z}\frac{z^{j-1}}{(j+1)!}\left(-2(j+1)z + j(j+1) + z^2\right). \end{aligned}$$

Set $q_1'(z) = 0$ and note that $q_1(0) = \lim_{z\to\infty} q_1(z) = 0$. Hence, the maximum of $|q_1(z)|$ is attained at either $z_1 := (j+1) - \sqrt{j+1}$ or $z_2 := (j+1) + \sqrt{j+1}$. We first consider the function value at z_1:

$$\begin{aligned} |q_1(z_1)| &= e^{-z_1}\frac{z_1^{j+1}}{(j+1)!}\left|\frac{j+1}{z_1} - 1\right| \\ &\leq e^{-(j+1)+\sqrt{j+1}}((j+1) - \sqrt{j+1})^{j+1}\frac{e^{j+1}}{\sqrt{2\pi}(j+1)^{j+1+1/2}}\frac{1}{\sqrt{j+1}-1} \\ &\leq e^{\sqrt{j+1}}\left(1 - \frac{1}{\sqrt{j+1}}\right)^{j+1}\frac{1}{\sqrt{2\pi}\sqrt{j+1}}\frac{1}{\sqrt{j+1}-1} \\ &\leq \frac{1}{\sqrt{2\pi}((j+1) - \sqrt{j+1})}. \end{aligned}$$

By the same reasoning, we also have $|q_1(z_2)| \leq 1/(\sqrt{2\pi}((j+1) + \sqrt{j+1}))$ for z_2. Analogously, to establish the second inequality, we first denote

$$q_2(z) := e^{-z} \left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!} \right).$$

Then, the derivative of $q_2(z)$ is

$$q_2'(z) = e^{-z}\frac{z^j}{(j+1)!}\left(-(2j+3)z + (j+1)^2 + z^2\right).$$

Set $q_2'(z) = 0$ and note that $q_2(0) = \lim_{z\to\infty} q_2(z) = 0$. Hence, the maximum of $|q_2(z)|$ is attained at either $z_3 := ((2j+3) - \sqrt{4j+5})/2$ or $z_4 := ((2j+3) + \sqrt{4j+5})/2$. Furthermore, note that both $|z_3|, |z_4| \le 2(j+2)$. Therefore, we obtain

$$|q_2(z_3)| = |z_3||q_1(z_3)| \le 2(j+2)\max_z |q_1(z)| \le \frac{2(j+2)}{\sqrt{2\pi}((j+1) - \sqrt{j+1})} \le 5, \forall j \ge 1.$$

Finally, the same proof also shows that $|q_2(z_4)| \le 5$. □

4.I.1 Proving Theorem 2: The L_1 Distance

Now, let us focus on the problem of estimating the L_1 distance between the unknown distribution $p \in \Delta_k$ and a given distribution $q \in \Delta_k$. Since our estimator is constructed symbol by symbol, it suffices to consider the problem of approximating $\ell_q(x) := |x - q| - q$.

Let $g_{n+1}(j) := (n+1)\ell_q\left(\frac{j}{n+1}\right)$. We note that $r_{na}(j)$ equals 0 for all but at most two different values of j. Therefore, by Lemma 19, for all $z \in I_n'$, we have $|t_{na}''(z)| \lesssim 1$, and $|t_{na}''(z)| \lesssim z^{-1}$, where the first and second inequalities resemble Property 3 and 4 in Section 4.H.2, respectively. Using arguments similar to those in Section 4.H.2 and 4.H.3, we can construct an estimator for $D_q(p)$ that provides the guarantees stated in Theorem 2. Note that concavity/convexity is actually not crucial for establishing the final result in Section 4.H.2. Also note that we need to replace our analysis in Section 4.E.2 and 4.F.2 for the corresponding large-probability estimator by that in [HOSW18].

4.I.2 Proving Theorem 3: General Additive Properties

More generally, our result on L_1 distance extends to any additive property $F(p) = \sum_{i\in[k]} f_i(p_i)$ that satisfies the simple condition: f_i is $O(1)$-Lipschitz, for all i. Without loss of generality, assume that all functions f_i's are 1-Lipschitz and satisfy $f_i(0) = 0$. By the previous derivations, we immediately have $|t''_{na}(z)| \le 6$, which recovers Property 3 in Section 4.H.2. Again, concavity/convexity is actually unnecessary for establishing the final result in Section 4.H.2. The proof will be complete if we also recover Property 4 in that section. In other words, we only need to show $|t''_{na}(z)z| \lesssim 1$, where

$$t''_{na}(z)z = e^{-z}\sum_{j=0}^{\infty} r_{na}(j+1)\left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!}\right) - e^{-z}z\cdot r_{na}(0).$$

Fix $z \in I'_n$ and treat it as a constant. Let $b_j := r_{na}(j+1)$ and $a_j := e^{-z}\left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!}\right)$. By Lemma 19, we have $|a_j| \le 5, \forall j \ge 1$. Note that there is need to worry about the slack term $e^{-z}zr_{na}(0)$ and the first term in the sum which corresponds to $j = 0$, because both terms contribute at most $O(1)$ in absolute value to the above expression for any $z \ge 0$. The key observation is that any consecutive partial sum of sequence $\{b_j\}_{j\ge1}$ is also bounded by $O(1)$ in magnitude. Specifically, for any $n_1, n_2 \in \mathbb{Z}^+$ satisfying the inequality $n_1 + 2 \le n_2$,

$$\begin{aligned}
\left|\sum_{j=n_1}^{n_2} b_j\right| &= \left|\sum_{j=n_1}^{n_2} r_{na}(j+1)\right| \\
&= \left|\sum_{j=n_1}^{n_2} (g_{na}(j+3) + g_{na}(j+1) - 2g_{na}(j+2))\right| \\
&= \left|\sum_{j=n_1+3}^{n_2+3} g_{na}(j) + \sum_{j=n_1+1}^{n_2+1} g_{na}(j) - 2\sum_{j=n_1+2}^{n_2+2} g_{na}(j)\right| \\
&= |(g_{na}(n_2+3) - g_{na}(n_2+2)) + (g_{na}(n_1+1) - g_{na}(n_1+2))| \\
&\le 2.
\end{aligned}$$

Furthermore, the sequence $\{a_j\}_{j\geq 1}$ can change its monotonicity at most two times, which can be proved by considering the sign of $a_j - a_{j-1}$. More concretely,

$$\begin{aligned}\operatorname{sign}\left(a_j - a_{j-1}\right) &= \operatorname{sign}\left(e^{-z}\left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!}\right) - e^{-z}\left(\frac{z^{j}}{(j-1)!} - \frac{z^{j+1}}{j!}\right)\right) \\ &= \operatorname{sign}\left(2(j+1)z - z^2 - (j+1)j\right) \\ &= \operatorname{sign}\left(-j^2 + j(2z-1) + (2z - z^2)\right).\end{aligned}$$

Since z is fixed, the last expression can change its value at most two times as j increases from 0 to infinity. The last piece of the proof is the following corollary of the well-known Abel's inequality.

Lemma 20. *Let $\{a'_j\}_{j=1}^m$ be a sequence of real numbers that is either increasing or decreasing, and let $\{b'_j\}_{j=1}^m$ be a sequence of real or complex numbers. Then, for $B'_t := \sum_{j=1}^t b'_t$,*

$$|\sum_{j=1}^{m} a'_j b'_j| \leq \max_{t=1,\ldots,m} |B'_t|(2|a'_n| + |a'_1|).$$

By the previous discussions, we can find two indices j_1 and j_2, such that $\{a_j\}_{j=1}^{j_1}$, $\{a_j\}_{j=j_1+1}^{j_2}$, and $\{a_j\}_{j\geq j_2+1}$ are all monotone subsequences.

Then, we apply Lemma 20 to each subsequence and further bound the resulting quantity by the inequalities established above: $\left|\sum_{j=n_1}^{n_2} b_j\right| \lesssim 1$ and $|a_j| \leq 6, \forall j \geq 1$. This concludes the proof.

Finally, we point out that the above argument applies to a much broader class of additive properties beyond the Lipschitz ones, which is not addressed here for the sake of clarity and simplicity.

4.J A Competitive Estimator for Support Size

4.J.1 Estimator Construction

Denote by p and S_p an unknown distribution and its support size, respectively. For $\varepsilon \leq e^{-2}$, redefine the amplification parameter as $a := |\log^{-2}\varepsilon| \cdot \log S_p$. Let X^{na} be an i.i.d. sample sequence drawn from p, and N_i'' be the number of times symbol i appears empirically.

The na-sample empirical estimator approximates the support size $S_p = \sum_{i\in[k]} \mathbb{1}_{p_i>0}$ by

$$\hat{S}^E(X^{na}) := \sum_{i\in[k]} \mathbb{1}_{N_i''>0}.$$

Taking expectation, we have

$$\mathbb{E}[\hat{S}^E(X^{na})] := \sum_{i\in[k]} \mathbb{E}[\mathbb{1}_{N_i''>0}] = \sum_{i\in[k]} (1-(1-p_i)^{na}).$$

For a length-$\mathrm{Poi}(n)$ sample sequence X^N, denote by ϕ_j the number of symbols that appear j times. Following [ADOS17, OSW16], we can estimate $\mathbb{E}[\hat{S}^E(X^{na})]$ by

$$\hat{S}(X^N) := \sum_{j=1}^{\infty} \phi_j (1-(-(a-1))^j \Pr(Z \geq j)),$$

where $Z \sim \mathrm{Poi}(r)$ for some *smoothing parameter* r. Similar to the previous notation, we define N_i as the number of times symbol i appears in X^N. Then, all the N_i's are mutually independent.

4.J.2 Bounding the Bias

The following lemma bounds the bias of $\hat{S}(X^N)$ in estimating $\mathbb{E}[\hat{S}^E(X^{na})]$.

Lemma 21. *For any $a \geq 1$,*

$$|\mathbb{E}[\hat{S}(X^N)] - \mathbb{E}[\hat{S}^E(X^{na})]| \leq \min\left\{na, S_p\right\} e^{-r} + 2.$$

Proof. Note that for any $m \geq 0$ and $p \in [0,1]$,

$$0 \leq e^{-mp} - (1-p)^m \leq 2p.$$

Hence, we obtain

$$\begin{aligned}
&|\mathbb{E}[\hat{S}(X^N)] - \mathbb{E}[\hat{S}^E(X^{na})]| \\
&= \left| \mathbb{E}\left[\sum_j^\infty \phi_j\right] - \mathbb{E}\left[\sum_j^\infty \phi_j(-(a-1))^j \Pr(Z \geq j)\right] - \sum_{i\in[k]} (1-(1-p_i)^{na}) \right| \\
&= \left| \sum_{i\in[k]} (1-e^{-np_i}) - \mathbb{E}\left[\sum_j^\infty \phi_j(-(a-1))^j \Pr(Z \geq j)\right] - \sum_{i\in[k]} (1-(1-p_i)^{na}) \right| \\
&\leq \left| \sum_{i\in[k]} (-e^{-np_i}) - \mathbb{E}\left[\sum_j^\infty \phi_j(-(a-1))^j \Pr(Z \geq j)\right] - \sum_{i\in[k]} (-e^{-nap_i}) \right| + 2\sum_{i\in[k]} p_i \\
&= \left| \sum_{i\in[k]} e^{-np_i}(e^{-n(a-1)p_i} - 1) - \mathbb{E}\left[\sum_j^\infty \phi_j(-(a-1))^j \Pr(Z \geq j)\right] \right| + 2 \\
&\leq \min\left\{na, S_p\right\} e^{-r} + 2,
\end{aligned}$$

where the last step follows by Lemma 7 and Corollary 2 in [OSW16]. □

4.J.3 Bounding the Mean Absolute Deviation

Bounds for $\hat{S}(X^N)$

In this section, we analyze the mean absolute deviation of $\hat{S}(X^N)$. To do this, we need the following two lemmas. The first lemma bounds the coefficients of this estimator.

Lemma 22 ([ADOS17])**.** *For any $j \geq 1$ and $a \geq 1$,*

$$|1-(-(a-1))^j \Pr(Z \geq j)| \leq 1 + e^{r(a-1)}.$$

The second lemma is the well-known McDiarmid's inequality.

Lemma 23. *Let $Y_1, \ldots, Y_m$ be independent variables taking values in ranges $R_1, \ldots, R_m$, and let $F : R_1 \times \ldots \times R_m \to C$ with the property that if one freezes all but the w^{th} coordinate of $F(y_1, \ldots, y_m)$ for some $1 \leq w \leq m$, then F fluctuates by only most $c_w > 0$, thus*

$$|F(y_1, \ldots, y_{w-1}, y_w, y_{w+1}, \ldots, y_m) - F(y_1, \ldots, y_{w-1}, y'_w, y_{w+1}, \ldots, y_m)| \leq c_w$$

for all $y_j \in R_j$ and $y'_w \in R_w$ for $1 \leq j \leq m$. Then for any $\lambda > 0$, one has $\Pr(|F(Y) - \mathbb{E}[F(Y)]| \geq \lambda\sigma) \leq C\exp(-c\lambda^2)$ for some absolute constants $C, c > 0$, where $\sigma^2 := \sum_{j=1}^m c_j^2$.

Note that $\hat{S}(X^N)$, viewed as a function of N_i's with indexes i satisfying $p_i \neq 0$, fulfills the conditions described in Lemma 23, with parameter $m = S_p$ and $c_w = 2 + 2e^{r(a-1)}$ for all $1 \leq w \leq m$. Therefore, for $\sigma^2 := 4S_p(1+e^{r(a-1)})^2$,

$$\Pr(|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]| \geq \lambda\sigma) \leq C\exp(-c\lambda^2).$$

This inequality further implies

$$\begin{aligned}
\mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]\right| &= \int_0^\infty \Pr(|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]| \geq t)\, dt \\
&= \sigma \int_0^\infty \Pr(|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]| \geq \lambda\sigma)\, d\lambda \\
&\leq C\sigma \int_0^\infty \exp(-c\lambda^2) d\lambda \\
&\lesssim \sqrt{S_p}(1+e^{r(a-1)}).
\end{aligned}$$

Analogously, treating $\hat{S}(X^N)$ as a function of X_i's yields

$$\mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]\right| \lesssim \sqrt{n}(1+e^{r(a-1)}).$$

Consolidating the previous results, we obtain

$$\mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]\right| \lesssim \sqrt{\min\left\{S_p, n\right\}}(1+e^{r(a-1)}).$$

Bounds for $\hat{S}^E(X^{na})$

The following lemma bounds the variance of $\hat{S}^E(X^{na})$ in terms of S_p.

Lemma 24. *For $m \geq 1$ and $X^m \sim p$,*

$$\mathrm{Var}(\hat{S}^E(X^m)) \lesssim S_p.$$

Proof. In this proof, we slightly abuse the notation and denote by N_i the number of times symbol i appears in X^m. Incorporating the definition,

$$\mathrm{Var}(\hat{S}^E(X^m)) = \mathrm{Var}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right) = \mathbb{E}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right)^2 - \left(\mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right]\right)^2.$$

Let Y^M be an independent length-$\mathrm{Poi}(m)$ sample sequence from p, and N_i' be the number of times

symbol i appearing in X^M. Then,

$$\begin{aligned}
\mathbb{E}\left(\sum_{i:p_i>0}\mathbb{1}_{N_i>0}\right)^2 &= \mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i>0}+\sum_{i\neq j:p_i>0,p_j>0}\mathbb{1}_{N_i>0}\mathbb{1}_{N_j>0}\right] \\
&= \sum_{i:p_i>0}(1-\mathbb{E}[\mathbb{1}_{N_i=0}])+\sum_{i\neq j:p_i>0,p_j>0}\mathbb{E}[(1-\mathbb{1}_{N_i=0})(1-\mathbb{1}_{N_j=0})] \\
&= \sum_{i:p_i>0}(1-(1-p_i)^m) \\
&+ \sum_{i\neq j:p_i>0,p_j>0}\left(1-(1-p_i)^m-(1-p_j)^m+(1-p_i-p_j)^m\right).
\end{aligned}$$

Note that for any $m \geq 0$ and $p \in [0,1]$,

$$0 \leq e^{-mp}-(1-p)^m \leq 2p.$$

Then, we must have both

$$|(1-(1-p_i)^m)-(1-e^{-mp_i})| \leq 2p_i$$

and

$$|(1-(1-p_i)^m-(1-p_j)^m+(1-p_i-p_j)^m)-(1-e^{-mp_i}-e^{-mp_j}+e^{-m(p_i+p_j)})| \leq 4(p_i+p_j).$$

Therefore,

$$\begin{aligned}
\left|\mathbb{E}\left(\sum_{i:p_i>0}\mathbb{1}_{N_i>0}\right)^2-\mathbb{E}\left(\sum_{i:p_i>0}\mathbb{1}_{N_i'>0}\right)^2\right| &\leq \sum_{i:p_i>0}2p_i+\sum_{i\neq j:p_i>0,p_j>0}4(p_i+p_j) \\
&\leq 4\sum_{i:p_i>0}\sum_{j:p_j>0}(p_i+p_j) \\
&\leq 8S_p.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left|\left(\mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i>0}\right]\right)^2-\left(\mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i'>0}\right]\right)^2\right| \\
&=\left|\mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i>0}\right]-\mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i'>0}\right]\right|\left|\mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i>0}\right]+\mathbb{E}\left[\sum_{i:p_i>0}\mathbb{1}_{N_i'>0}\right]\right| \\
&\leq\Big(\sum_{i:p_i>0}2p_i\Big)\cdot 2S_p \\
&\leq 4S_p.
\end{aligned}$$

Finally, note that changing the value of a single observation changes the value of $\sum_{i:p_i>0}\mathbb{1}_{N_i'>0}$ by at most one. Hence, by McDiarmid's inequality,

$$\mathrm{Var}\left(\sum_{i:p_i>0}\mathbb{1}_{N_i'>0}\right)\lesssim S_p.$$

The triangle inequality combines the previous inequalities and yields

$$\mathrm{Var}\left(\sum_{i:p_i>0}\mathbb{1}_{N_i>0}\right)\lesssim S_p. \qquad \square$$

By Jensen's inequality, the above lemma implies that

$$\mathbb{E}\left|\hat{S}^E(X^{na})-\mathbb{E}[\hat{S}^E(X^{na})]\right|\leq\sqrt{\mathrm{Var}(\hat{S}^E(X^{na}))}\lesssim\sqrt{S_p}.$$

4.J.4 Proving Theorem 4

Setting $r=|\log\varepsilon|$, we obtain

$$e^{r(a-1)}\leq S_p^{|\log^{-1}\varepsilon|}$$

and

$$e^{-r} = e^{-|\log \varepsilon|} = \varepsilon.$$

Therefore, by the previous results,

$$\begin{aligned}\mathbb{E}\left|\hat{S}(X^N) - \hat{S}^E(X^{na})\right| &\leq \mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}^E(X^{na})]\right| + \mathbb{E}\left|\mathbb{E}[\hat{S}^E(X^{na})] - \hat{S}^E(X^{na})\right| \\ &\lesssim S_p^{|\log^{-1}\varepsilon|+\frac{1}{2}} + S_p \cdot \varepsilon.\end{aligned}$$

Normalize both sides by S_p. Then,

$$\mathbb{E}\left|\frac{\hat{S}(X^N)}{S_p} - \frac{\hat{S}^E(X^{na})}{S_p}\right| \lesssim S_p^{|\log^{-1}\varepsilon|-\frac{1}{2}} + \varepsilon.$$

4.K A Competitive Estimator for Support Coverage

4.K.1 Estimator Construction

Recall that $c(p) = 1-(1-p_i)^m$, where m is a *given parameter*. For $\varepsilon \leq e^{-2}$, redefine the amplification parameter as $a := |\log^{-2}\varepsilon| \cdot \log C_p$. Similar to the last section, let X^{na} be an independent length-na sample sequence drawn from p, and N_i'' be the number of times symbol i appears empirically.

The na-sample empirical estimator estimates the m-sample support coverage $C_p = \sum_{i\in[k]} c(p_i)$ by

$$\hat{C}^E(X^{na}) := \sum_{i\in[k]} c\left(\frac{N_i''}{na}\right) = \sum_{i\in[k]} \left(1-\left(1-\frac{N_i''}{na}\right)^m\right).$$

Taking expectation, we obtain

$$\mathbb{E}[\hat{C}^E(X^{na})] = \sum_{i\in[k]} \mathbb{E}\left[1-\left(1-\frac{N_i''}{na}\right)^m\right].$$

For the ease of exposition, let us denote

$$T(p) := \sum_{i\in[k]} \mathbb{E}\left[1 - e^{-m\frac{N_i''}{na}}\right].$$

Noting that for any $t \geq 1$ and $p \in [0,1]$,

$$|e^{-tp} - (1-p)^t| \leq 2p,$$

hence, we have

$$|\mathbb{E}[\hat{C}^E(X^{na})] - T(p)| \leq \sum_{i\in[k]} \mathbb{E}\left[2 \cdot \frac{N_i''}{na}\right] = 2.$$

Then, it suffices to estimate $T(p)$, which satisfies

$$\begin{aligned}
T(p) &= \sum_{i\in[k]} \left(1 - \mathbb{E}\left[e^{-m\frac{N_i''}{na}}\right]\right) \\
&= \sum_{i\in[k]} \left(1 - \sum_{j=0}^{na} \binom{na}{j} p_i^j (1-p_i)^{na-j} e^{-m\frac{j}{na}}\right) \\
&= \sum_{i\in[k]} \left(1 - \sum_{j=0}^{na} \binom{na}{j} \left(p_i \cdot e^{-\frac{m}{na}}\right)^j (1-p_i)^{na-j}\right) \\
&= \sum_{i\in[k]} \left(1 - \left(1 - p_i(1 - e^{-\frac{m}{na}})\right)^{na}\right).
\end{aligned}$$

Analogous to the definition of $T(p)$, let us denote

$$T_1(p) := \sum_{i\in[k]} \left(1 - \exp\left(-na(1 - e^{-\frac{m}{na}})p_i\right)\right).$$

Since $(1 - e^{-\frac{m}{na}}) \cdot p_i \in [0,1]$, we must have

$$|T(p) - T_1(p)| \leq \sum_{i\in[k]} 2(1 - e^{-\frac{m}{na}})p_i \leq 2.$$

Define a new amplification parameter $a' := a(1-e^{-\frac{m}{na}})$. Then, we can express $T_1(p)$ as

$$T_1(p) := \sum_{i\in[k]} \left(1-\exp\left(-na'p_i\right)\right).$$

For simplicity, we will assume that $m \geq 1.5n$ and $a > 1.8$, ensuring

$$a' = a(1-e^{-\frac{m}{na}}) \geq a(1-e^{-\frac{1.5}{a}}) > 1.$$

Analogous to case of support size estimation, we draw a length-Poi(n) sample sequence X^N and estimate $\mathbb{E}[\hat{C}^E(X^{na})]$ by the estimator

$$\hat{C}(X^N) := \sum_{j=1}^{\infty} \phi_j (1-(-(a'-1))^j \Pr(\text{Poi}(r) \geq j)),$$

where ϕ_j denotes the number of symbols appearing j times.

4.K.2 Bounding the Bias

We bound the bias of $\hat{C}(X^N)$ in estimating $\mathbb{E}[\hat{C}^E(X^{na})]$ as follows.

$$\begin{aligned}
|\mathbb{E}[\hat{C}(X^N)] - \mathbb{E}[\hat{C}^E(X^{na})]| &\leq |\mathbb{E}[\hat{C}(X^N)] - T_1(p)| + |T_1(p) - \mathbb{E}[\hat{C}^E(X^{na})]| \\
&\leq |\mathbb{E}[\hat{C}(X^N)] - T_1(p)| + 4 \\
&= \left| \sum_{i\in[k]} e^{-np_i}(e^{-n(a'-1)p_i} - 1) \right. \\
&\quad \left. - \sum_{i\in[k]} e^{-np_i} \sum_{j=1}^{\infty} \frac{(-(a'-1)np_i)^j}{j!} \Pr(\text{Poi}(r) \geq j) \right| + 4 \\
&\leq \left| \sum_{i\in[k]} e^{-np_i} \left(\sum_{j=1}^{\infty} \frac{(-(a'-1)np_i)^j}{j!} \Pr(\text{Poi}(r) < j) \right) \right| + 4.
\end{aligned}$$

To bound the last sum, we need the following lemma.

Lemma 25. *For any $y, r \geq 0$,*

$$\left| \sum_{j=1}^{\infty} \frac{(-y)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right| \leq e^{-r}(1 - e^{-y}).$$

Proof. By Lemma 6 of [OSW16],

$$\begin{aligned}
\left| \sum_{j=1}^{\infty} \frac{(-y)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right| &\leq \max_{s \leq y} \left| \mathbb{E}_{L \sim \mathrm{Poi}(r)} \left[\frac{(-s)^L}{L!} \right] \right| (1 - e^{-y}) \\
&= \max_{s \leq y} \left| J_0(2\sqrt{sr}) \right| e^{-r}(1 - e^{-y}) \\
&\leq e^{-r}(1 - e^{-y}),
\end{aligned}$$

where J_0 is the first-order Bessel function of the first kind, and satisfies the elegant inequality $|J_0(x)| \leq 1, \forall x \geq 0$ [AS65]. □

Leveraging the above lemma, we obtain

$$\begin{aligned}
|\mathbb{E}[\hat{C}(X^N)] - \mathbb{E}[\hat{C}^E(X^{na})]| &\leq \left| \sum_{i \in [k]} e^{-np_i} \left(\sum_{j=1}^{\infty} \frac{(-(a'-1)np_i)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right) \right| + 4 \\
&\leq e^{-r} \sum_{i \in [k]} e^{-np_i}(1 - e^{-(a'-1)np_i}) + 4 \\
&\leq e^{-r} \sum_{i \in [k]} (1 - e^{-na'p_i}) + 4.
\end{aligned}$$

Note that $na' = na(1 - e^{-\frac{m}{na}}) \leq m$. Therefore,

$$|\mathbb{E}[\hat{C}(X^N)] - \mathbb{E}[\hat{C}^E(X^{na})]| \leq e^{-r} \sum_{i \in [k]} (1 - e^{-mp_i}) + 4 = e^{-r} C_p + 4.$$

4.K.3 Bounding the Mean Absolute Deviation

Bounds for $\hat{C}(X^N)$

First, we bound the mean absolute deviation of $\hat{C}(X^N)$ in terms of C_p. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left|\hat{C}(X^N) - \mathbb{E}[\hat{C}(X^N)]\right| &\leq \sqrt{\mathrm{Var}\left(\hat{C}(X^N)\right)} \\
&= \sqrt{\sum_{i\in k} \mathrm{Var}\left(\sum_{j=1}^{\infty} \mathbb{1}_{N_i=j}(1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\geq j))\right)} \\
&\leq \sqrt{\sum_{i\in k} \mathbb{E}\left[\left(\sum_{j=1}^{\infty} \mathbb{1}_{N_i=j}(1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\geq j))\right)^2\right]} \\
&= \sqrt{\sum_{i\in k}\sum_{j=1}^{\infty} \mathbb{E}\left[\mathbb{1}_{N_i=j}\right](1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\geq j))^2} \\
&\leq (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-e^{-np_i})}.
\end{aligned}$$

By our assumption that $m \geq 1.5n$,

$$\begin{aligned}
\mathbb{E}[|\hat{C}(X^N) - \mathbb{E}[\hat{C}(X^N)]|] &\leq (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-e^{-np_i})} \\
&\leq (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-e^{-mp_i})} \\
&\leq (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-(1-p_i)^m)} \\
&= (1+e^{r(a'-1)})\sqrt{C_p}.
\end{aligned}$$

Bounds for $\hat{C}^E(X^{na})$

Next, we bound the mean absolute deviation of the na-sample empirical estimator. To deal with the dependence among the counts N_i'''s, we need the following lemma [JDP83].

Definition 1. *Random variables $X_1,\ldots,X_S$ are said to be* negatively associated *if for any pair of disjoint subsets A_1,A_2 of $1,2,\ldots,S$, and any component-wise increasing functions f_1,f_2,*

$$Cov(f_1(X_i,i\in A_1),f_2(X_j,j\in A_2))\leq 0.$$

The next result can be used to check whether random variables are negatively associated.

Lemma 26. *Let $X_1,\ldots,X_S$ be S independent random variables with log-concave densities. Then the joint conditional distribution of $X_1,\ldots,X_S$ given $\sum_{i=1}^{S}X_i$ is negatively associated.*

Lemma 26 shows that N_i'''s are negatively correlated. Furthermore, note that

$$c^*(x):=1-\left(1-\frac{x}{na}\right)^m$$

is an increasing function, and we can write the quantity of interest as

$$\hat{C}^E(X^{na}):=\sum_{i\in[k]}c^*(N_i'').$$

Hence, for any $i,j\in[k]$ such that $i\neq j$, $\mathrm{Cov}(c^*(N_i''),c^*(N_j''))\leq 0$. Consequently,

$$\begin{aligned}
\mathrm{Var}\left(\hat{C}^E(X^{na})\right) &= \sum_{i\in[k]}\mathrm{Var}(c^*(N_i''))+2\sum_{i,j\in[k],i\neq j}\mathrm{Cov}(c^*(N_i''),c^*(N_j'')) \\
&\leq \sum_{i\in[k]}\mathrm{Var}(c^*(N_i''))\leq\sum_{i\in[k]}\mathbb{E}(c^*(N_i''))^2 \\
&= \sum_{i\in[k]}\mathbb{E}\left[\sum_{j=0}^{na}\mathbb{1}_{N_i=j}(C^*(j))^2\right]\leq\sum_{i\in[k]}\sum_{j=1}^{na}\mathbb{E}\left[\mathbb{1}_{N_i=j}\right] \\
&= \sum_{i\in[k]}(1-(1-p_i)^{na}).
\end{aligned}$$

Without loss of generality, we will assume that a is a positive integer. Then,

$$\begin{aligned}\sum_{i\in[k]}(1-(1-p_i)^{na}) &= \sum_{i\in[k]}(1-(1-p_i)^{n})(\sum_{j=0}^{a-1}(1-p_i)^{nj}) \\ &\le a\sum_{i\in[k]}(1-(1-p_i)^{n}) \\ &\le a\sum_{i\in[k]}(1-(1-p_i)^{m}) \\ &= aC_p.\end{aligned}$$

Finally, Jensen's inequality implies

$$\mathbb{E}\left|\hat{C}^E(X^{na})-\mathbb{E}[\hat{C}^E(X^{na})]\right| \le \sqrt{\mathrm{Var}(\hat{C}^E(X^{na}))} \le \sqrt{aC_p}.$$

4.K.4 Proving Theorem 5

The triangle inequality consolidates the major inequalities in the previous sections and yields

$$\mathbb{E}\left|\hat{C}(X^N)-\hat{C}^E(X^{na})\right| \lesssim e^{-r}C_p+4+\sqrt{aC_p}+(1+e^{r(a'-1)})\sqrt{C_p}.$$

By the fact that $a' < a = |\log^{-2}\varepsilon|\cdot\log C_p$, we set $r = |\log\varepsilon|$ and obtain

$$\mathbb{E}\left|\hat{C}(X^N)-\hat{C}^E(X^{na})\right| \lesssim \varepsilon C_p+4+(1+C_p^{|\log^{-1}\varepsilon|}+\sqrt{\log C_p})\sqrt{C_p}.$$

Then, normalizing both sides by C_p gives

$$\mathbb{E}\left|\frac{\hat{C}(X^N)}{C_p}-\frac{\hat{C}^E(X^{na})}{C_p}\right| \lesssim C_p^{|\log^{-1}\varepsilon|-\frac{1}{2}}+\varepsilon.$$

Chapter 5

Profile Entropy: A Fundamental Measure for the Learnability and Compressibility of Distributions

5.1 Introduction

Recent research in statistical machine learning, ranging from neural-network training and online learning, to density estimation and property testing, has advanced evaluation criteria beyond worst-case analysis. New performance measures apply more refined metrics relating the algorithm's accuracy and efficiency to the problem's inherent structure.

Consider for example learning an unknown discrete distribution from its i.i.d.samples (see also Section 5.2.2). The classical worst-case analysis states that in the worst case, the number of samples required to estimate a distribution to a given KL-divergence grows linearly in the alphabet size.

However, this formulation is pessimistic, since distributions are rarely the worst possible, and many practical distributions can be estimated with significantly smaller samples. Furthermore,

once the sample is drawn, it reveals the distribution's complexity and hence the hardness of the learning task.

Going beyond worst-case analysis, one can design an *adaptive* learning algorithm whose theoretical guarantees vary according to the problem's simplicity. For example, [OS15] recently proposed an estimator that instance-by-instance achieves nearly the same performance as a genie algorithm designed with prior knowledge of the underlying distribution.

We introduce *profile entropy*, a fundamental measure for the complexity of discrete distributions, and show that it connects three vital scientific tasks: estimation, inference, and compression. The resulting algorithms have guarantees directly relating to the sample profile entropy, hence also adapt to the intrinsic simplicity of the tasks at hand.

The next subsection formalizes relevant concepts and useful notation.

Sample Profiles and Their Entropy

Consider an arbitrary sequence x^n over a finite or countably infinite alphabet $\mathcal{X}$. The *multiplicity* $\mu_y(x^n)$ of a symbol $y \in \mathcal{X}$ is the number of times y appears in x^n. The *prevalence* of an integer μ is the number $\varphi_\mu(x^n)$ of symbols in x^n with multiplicity μ. The *profile* of x^n is the multiset $\varphi(x^n)$ of multiplicities of the symbols in x^n. We refer to it as a profile of *length* n. For example, consider the sequence $x^7 = bananas$, in which a appears thrice, n appears twice, and b and s each appears once. Then, the profile of the sequence is multiset $\varphi(x^7) = \{3,2,1,1\}$.

The number $\mathcal{D}(S)$ of distinct elements in a multiset S is its *dimension*. For convenience, we also write $\mathcal{D}(x^n)$ for profile dimension. In the above example, we have $\mathcal{D}(x^7) = \mathcal{D}(\varphi(x^7)) = 3$, corresponding to values $1, 2$, and 3. The dimension of a length-n profile over $\mathcal{X}$ is at most $\min\{\sqrt{2n}, |\mathcal{X}|\}$. In general, the profile entropy $\mathcal{H}_n(p)$ is no more than $3\sqrt{n}$.

Let Δ be the collection of all discrete distributions, and $\Delta_\mathcal{X}$ be the collection of those over $\mathcal{X}$. Draw a size-n sample X^n from an arbitrary distribution in $p \in \Delta$. Then, the profile Φ^n of X^n is a random multiset whose distribution depends on only p and n. We therefore write $\Phi^n \sim p$, and

call $\mathcal{H}_n(p) := \mathrm{H}(\Phi^n)$ the *profile entropy* with respect to (p,n). For example, if we draw a sample of size $n = 3$ from $p = (\frac{1}{2}, \frac{1}{2})$, then profiles $\{1,1,1\}$, $\{2,1\}$, and $\{3\}$ appear with probabilities 0, $\frac{3}{4}$, and $\frac{1}{4}$, respectively. And the profile entropy is thus $\mathcal{H}_3(\frac{1}{2}, \frac{1}{2}) = H(0, \frac{3}{4}, \frac{1}{4}) \approx 0.56$.

Analogously, we call $\mathcal{D}_n := \mathcal{D}(\Phi^n)$, the *profile dimension* associated with (p,n), and write $\mathcal{D}_n \sim p$.

For notational simplicity, we will assume that $\mathcal{H}_n(p) \geq 1$ throughout the paper, and respectively write $a \simeq b$, $a \gtrsim b$, and $a \lesssim b$ instead of $a = \tilde{\Theta}(b)$, $a = \tilde{\Omega}(b)$, and $a = \tilde{O}(b)$, where the asymptotic notation hides logarithmic factors of n.

Applications of Sample Profiles

Sample profiles have essential applications in numerous aspects of scientific research, ranging from property inference to the study of degree distributions of networks/graphs.

Property inference As Section 5.2.3 shows, profiles are sufficient for inferring all symmetric properties, such as entropy, Rényi entropy, and support size, not only in the sense of sufficient statistics, but also in the sense of Theorem 3, stating that profile-based estimators are as good as any others.

Distribution learning The entropy of a sample profile, equaling its dimension in order with high probability (Theorem 1), directly characterizes how well we can estimate a distribution and approach the performance of the best human-designed estimator (Theorem 2), for every distribution.

Theory of long tail The notable long tail theory in economics [And06] describes the strategy of selling a large number of different items that each sells in relatively small quantities. The profile of the product selling data, and the induced (PML) probability multiset estimate (Section 5.2.3), accurately characterize the tail shape of the data, and that of the underlying distribution, respectively.

Password frequency lists In the research of password defense, it is vital to understand

the distribution of passwords. Due to security concerns, organizations typically do not publish the complete data displaying each password and its frequency. Instead, they reveal the anonymized list of password frequencies, with each password hashed or replaced by some dummy string, which is equivalent to showing the password data's profile.

Degree distributions of networks Degree distribution is one of the most widely studied attributes of networks (and graphs) that describes the fractions of nodes with different degrees. As the degree distribution ignores symbol labeling and focuses only on the frequency of each degree, it is equivalent to the profile of the node degree data.

5.2 Main Results

This paper aims to provide a thorough theory of profile entropy. Most of the results either are the first of their kind or significantly improve the state-of-the-art.

Specifically, Section 5.2.1 presents the fundamental equivalence relation between profile dimension and entropy (Thm. 1). Building on the equivalence, we respectively establish essential connections between profile entropy and the estimation of discrete distributions (Section 5.2.2; Thm. 2), inference of their properties (Section 5.2.3; Thm. 4), and compression of sample profiles (Section 5.2.4; Thm. 5). These results characterize how well one can compete with an instance-optimal algorithm for each task, over *every single distribution*. For a real sense of how profile entropy behaves, Section 5.2.5 ultimately determines its magnitude for three prominent structural distribution families, log-concave (Thm. 6), power-law (Thm. 7), and histogram (Thm. 8). Going even further, Section 5.3 presents several additional applications and extensions of our theory and results, including robust learning under domain symbol permutations, profile entropy for mixture models, competitive property estimation, adaptive testing and classification, and connection to the method of types.

For space considerations, we relegate detailed reviews on related work, most technical

proofs, and numerical experiments to the *supplementary material*.

5.2.1 Dimension-Entropy Equivalence of Profiles

The following theorem shows that for every distribution and sampling parameter n, the induced profile entropy and dimension are of the same order, with high probability.

Theorem 1 (Entropy-dimension equivalence)**.** *For any distribution $p \in \Delta$ and $\mathcal{D}_n \sim p$,*

$$\Pr(\mathcal{D}_n \simeq \mathcal{H}_n(p)) \geq 1 - \frac{1}{\sqrt{n}}.$$

We briefly comment on Theorem 1.

First, the theorem reveals a novel and fundamental relation between profile dimension and entropy. The relation also yields an intrinsic method to approximate the entropy of the sample's profile, a fairly involved functional, by only counting its dimension. In general, the number of possible length-n profiles of a distribution could be as large the number of partitions of integer n, and grows with n at a sub-exponential speed. Hence, even if p is known, computing the exact value of $\mathcal{H}_n(p)$ could be hard. On the other hand, if one applies our theorem to approximate $\mathcal{H}_n(p)$, we only need to draw a sample $X^n \sim p$, and find its profile dimension, which is computable in linear time through counting. Appendix 5.A.4 further illustrates how to estimate $\mathcal{H}_n$ with $m \ll n$ observations.

Second, the theorem serves as an essential building block for the subsequent results on distribution estimation, property inference, and profile compression, and enables us to establish their optimality. For example, in the process of deriving the optimal profile compression scheme and proving Theorem 5, we reason with $\mathcal{D}_n$ to bound the space of storing the profile, and utilize $\mathcal{H}_n(p)$ as an essential lower bound for lossless compression.

Third, despite the simple form of the theorem, the proof of this result is highly nontrivial, and relies on a recent breakthrough in solving the Shepp-Olkin monotonicity conjecture [HJ19],

which asserts that the entropy of a Poisson-binomial random variable is monotone in the defining success probabilities, over a hypercube near the origin.

5.2.2 Competitive (Instance-Optimal) Distribution Estimation

Estimating distributions from their samples is a statistical-inference cornerstone, and has numerous applications, ranging from biological studies [AIS+08] to language modeling [CG99]. A learning algorithm $\hat{p}$ in this setting is called a *distribution estimator*, which associates with every sequence x^n a distribution $\hat{p}(x^n) \in \Delta$. Given a sample $X^n \sim p$, we measure the performance of $\hat{p}$ in estimating distribution p by the Kullback-Leibler (KL) divergence $D(p \parallel \hat{p}(X^n))$.

Let $r_n(p, \hat{p}) := \min\{r : \Pr(D(p \parallel \hat{p}(X^n)) \leq r) \geq 9/10\}$ be the *minimal KL error* $\hat{p}$ could achieve with probability at least $9/10$. Then, the *worst-case error* of estimator $\hat{p}$ over $\mathrm{P} \subseteq \Delta$ is $r_n(\mathrm{P}, \hat{p}) := \max_{p \in \mathrm{P}} r_n(p, \hat{p})$, and the lowest worst-case error for P, achieved by the optimal estimator, is the *minimax error* $r_n(\mathrm{P}) := \min_{\hat{p}'} r_n(\mathrm{P}, \hat{p}')$. The most widely studied distribution set P is simply $\Delta_{\mathcal{X}}$. With $\mathcal{X}$ being finite, it has become a classical result that $r_n(\Delta_{\mathcal{X}}) = \Theta(|\mathcal{X}|/n)$, which is achievable, up to constant factors, by an add-constant estimator [BS04, KOPS15].

Beyond minimax Despite being minimax optimal, the $|\mathcal{X}|/n$-result and the algorithm, are not satisfiable from a practical point of view. The reason is that the formulation puts much of its emphasis on the worst-case performance, and ignores the intrinsic simplicity of p in a pessimistic fashion. Hence, the desire to design more efficient estimators for practical distributions, like power-law, or Poisson, has led to algorithms that possess adaptive estimation guarantees.

Concretely, the minimax formulation has two modifiable components – the collection P and the error function D. A common approach to specifying P is adding structural assumptions, such as monotonicity, m-modality, and log-concavity, which, in many cases, makes algorithm refinement possible by leveraging structural simplicity. An orthogonal approach to encouraging adaptability without imposing structures is to replace absolute error by relative error, which we illustrate below.

Competitive estimation Without strong prior knowledge on the underlying distribution, a reasonable estimator should *naturally* assign the same probability to symbols appearing an equal number of times. *Competitive estimation* calls for finding a universally near-optimal estimator that learns *every* distribution as well as the best natural estimator that knows the true distribution.

Denote by $\mathcal{N}$ the collection of all natural estimators. For any distribution $p \in \Delta$ and sample $X^n \sim p$, a given estimator $\hat{p}$ incurs, with respect to the best natural estimator knowing p, an instance-by-instance *relative KL error* of

$$\mathrm{D}_{\text{nat}}(p \parallel \hat{p}(X^n)) := D(p \parallel \hat{p}(X^n)) - \min_{\hat{q} \in \mathcal{N}} D(p \parallel \hat{q}(X^n)).$$

Analogous to the minimax formulation, we denote by

$$r_n^{\text{nat}}(p, \hat{p}) := \min\{r : \Pr(\mathrm{D}_{\text{nat}}(p \parallel \hat{p}(X^n)) \le r) \ge 9/10\}$$

the *minimal relative error* $\hat{p}$ achieves with probability at least $9/10$, by $r_n^{\text{nat}}(\mathrm{P}, \hat{p})$ the *worst-case relative erro*r of $\hat{p}$ over $\mathrm{P} \subseteq \Delta$, and by $r_n^{\text{nat}}(\mathrm{P})$ the *minimax relative error*.

Old and new results Initiating the competitive formulation, [OS15] show that a simple variant of the well-known Good-Turing estimator achieves $r_n^{\text{nat}}(\Delta) \lesssim 1/n^{1/3}$, and a more involved estimator in [AJOS13b] attains the optimal $r_n^{\text{nat}}(\Delta) \simeq 1/\sqrt{n}$. For a fully adaptive guarantee, [HO19b] further refine the bound and design an estimator $\hat{p}^\star$ achieving $r_n^{\text{nat}}(p, \hat{p}^\star) \lesssim \mathbb{E}_{\mathcal{D}_n \sim p}[\mathcal{D}_n/n] \lesssim r_n^{\text{nat}}(\Delta)$, for every $p \in \Delta$, but provide no lower bounds.

In this work, we completely characterize $r_n^{\text{nat}}(p, \cdot)$ with essentially matching lower and upper bounds. Surprisingly, we show that for nearly every sample size n, the quantity behaves like $\mathcal{H}_n(p)/n$.

Theorem 2 (Optimal competitive error)**.** *There is a near-linear-time computable estimator $\hat{p}^\star$, such that for any distribution p and n,*

$$r_n^{nat}(p, \hat{p}^\star) \lesssim \frac{\mathcal{H}_n(p)}{n},$$

where $\hat{p}^\star$ is the near linear-time computable estimator in [HO19b] mentioned above. On the other hand, for any $H \in [0, \sqrt{n})$,

$$\min_{\hat{p}} \max_{p:\mathcal{H}_n(p)\lesssim H} r_n^{nat}(p, \hat{p}) \gtrsim \frac{H}{n}.$$

First, we comment on the lower bound. Due to the classical minimax formulation, one might expect a lower bound in one of the following two forms – for every $\hat{p}$, $r_n^{\text{nat}}(p, \hat{p}) \gtrsim \mathcal{H}_n(p)/n$ for 1) some p or 2) every p. Form 1) turns out to be weak under the competitive formulation. Specifically, let p be a *trivial distribution* that assigns probability 1 to some symbol. Then, both the profile entropy and the error of the best natural estimator are zero, and the inequality trivially holds for every $\hat{p}$. Form 2), on the other hand, is purely impossible. Specifically, for every distribution p, one can set $\hat{p}$ to be best natural estimator, which leads to a relative error of zero, greater than $\mathcal{H}_n(p)/n$ unless p is trivial.

Second, we illustrate the significance of the result. The notable work of [HR18] shows that the number of integer partitions of n, which equals the number of length-n profiles, is at most $\exp(3\sqrt{n})$, implying that $\mathcal{H}_n(p) \le 3\sqrt{n}$ for any $p \in \Delta$. Therefore, the $\mathcal{H}_n(p)/n$ upper and lower bounds in the theorem yields $r_n^{\text{nat}}(\Delta) \simeq 1/\sqrt{n}$, recovering the main result of [OS15]. Besides set Δ, the theorem and its proof also imply nearly tight minimax relative-error bounds on numerous distribution sets P. Below, we present two results that fall into this category. In both cases, the minimax relative error is much lower than $1/\sqrt{n}$ if the parameter involved is $o(\sqrt{n})$.

The first example addresses set Δ_H of distributions whose n-sample profile entropy is H.

Corollary 1. *For any $H \gtrsim 1$, the minimax relative error over Δ_H is $r_n^{nat}(\Delta_H) \simeq H/n$.*

For a more concrete example, denote by $\mathcal{L}_\sigma$ the collection of log-concave distributions over $\mathbb{Z}$ whose variance is σ^2. Theorem 2 and the profile entropy bounds in Theorem 6 imply

Corollary 2. *For any* $1 \lesssim \sigma \leq \sqrt{n}$*, the minimax relative error over* $\mathcal{L}_\sigma$ *is* $r_n^{nat}(\mathcal{L}_\sigma) \simeq \sigma/n$.

5.2.3 Competitive-Optimal Property Inference

Numerous practical applications call for inferring *property values* of an unknown distribution from its samples, including entropy for graphical modeling [KF09], Rényi entropy for sequential decoding [Ari96], and support size for species richness estimation [Mag13]. Therefore, *property inference* has attracted considerable attention over the past few decades. For interested readers, please refer to Appendix 5.B.3 for a detailed two-page review of prior works and discussions about relevant methods.

Property inference Formally, a *distribution property* over some collection $\mathrm{P} \subseteq \Delta$ is a functional $f : \mathrm{P} \to \mathbb{R}$ that associates with each distribution a real value. Given a sample X^n from an unknown distribution $p \in \mathrm{P}$, the problem of interest is to infer the value of $f(p)$. For this purpose, we employ another functional $\hat{f} : \mathcal{X}^* \to \mathbb{R}$, an *estimator* mapping every sample to a real value. We measure the statistical efficiency of $\hat{f}$ in approximating f over P by its *absolute error* $|\hat{f}(X^n) - f(p)|$.

Given $X^n \sim p \in \mathrm{P}$, the *minimal absolute error rate*, or simply *error*, that $\hat{f}$ achieves with probability at least $9/10$ is $r_n(p, \hat{f}) := \min\{r : \Pr(|\hat{f}(X^n) - f(p)| \leq r) \geq 9/10\}$, where the dependence on f is *implicit*. While p is often unknown, the *worst-case error* of an estimator $\hat{f}$ over all distributions in P is $r_n(\mathrm{P}, \hat{f}) := \max_{p \in \mathrm{P}} r_n(p, \hat{f})$, and the lowest worst-case error for P, achieved by the optimal estimator, is the *minimax error* $r_n(\mathrm{P}) := \min_{\hat{f}'} r_n(\mathrm{P}, \hat{f}')$.

Profile maximum likelihood An important class of properties is the collection of symmetric ones, which encompasses numerous well-known distribution characteristics, such as Shannon entropy, Rényi entropy, support size, and ℓ_1 distance to the uniform distribution. Symmetry connects the estimation of such property to the sample profile, a sufficient statistic for the task in hand. The general principle of maximum likelihood then provides an intuitive estimator, *profile maximum likelihood (PML)* [OSVZ04], that maximizes the probability of observing the

profile.

Naturally and generally, we study symmetric property inference over a distribution collection $\mathrm{P} \subseteq \Delta$ that is also *symmetric*, i.e., if $p \in \mathrm{P}$, then P as well contains all the symbol-permuted versions of p. For every sample $x^n \in \mathcal{X}^n$ and symmetric P, the *PML estimator* over P maps x^n to a distribution

$$\mathcal{P}_\varphi(x^n) := \arg\max_{p \in \mathrm{P}} \Pr_{X^n \sim p}(\varphi(X^n) = \varphi(x^n)).$$

Given a sample $X^n \sim p \in \mathrm{P}$ and a symmetric property p, the PML plug-in estimator uses $f \circ \mathcal{P}(X^n)$ to estimate $f(p)$. The PML estimator often behaves differently from the classical empirical distribution estimator. For example, if $P = \Delta$ and $\varphi = \{2,1,1\}$, the PML estimate turns out to be $\mathcal{P}_\varphi = (\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$, deviating from the empirical distribution $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ by 0.8 in (sorted) L_1 distance.

Recent researches [ADOS17, HO19a] show that for an extensive family of symmetric properties, including the previously mentioned four, the PML plug-in estimator *universally* achieves minimax error in the large-alphabet regime, up to constant factors.

The formulation of PML makes it part of two estimator classes, the maximum-likelihood and the *profile-based*, where the latter corresponds to estimators whose values depend on only the profile. The theorem below shows that profile-based estimators are sufficient for inferring symmetric properties.

Theorem 3 (Sufficiency of profiles)**.** *For any symmetric property f and set $\mathrm{P} \subseteq \Delta$, and estimator $\hat{f}$, we can construct an explicit estimator $\hat{F}$ over length-n profiles satisfying*

$$r_n(p, \hat{f}) = r_n(\mathrm{P}, \hat{F} \circ \varphi),$$

where both estimators can have independent randomness.

The next result shows that the PML estimator is adaptive to the simplicity of underlying distributions in inferring all symmetric properties, over any symmetric P. Specifically, the theorem

states that the n-sample PML plug-in essentially performs as well as the optimal $n/\mathcal{H}_n(p)$-sample estimator, which approaches the performance of the optimal n-sample estimator if p has a small $\mathcal{H}_n(p)$. Furthermore, for any property and estimator, there is a symmetric set P' for which this $1/\mathcal{H}_n(p)$ ratio is *optimal*.

Theorem 4 (Competitiveness of PML)**.** *For any symmetric property f and set $\mathrm{P} \subseteq \Delta$, and every distribution $p \in \mathrm{P}$, the PML plug-in estimator satisfies*

$$r_n\big(p, f \circ \mathcal{P}_\varphi\big) \le 2r_{n_p}(\mathrm{P}),$$

where $n_p := \simeq n/\mathcal{H}_n(p)$. On the other hand, for any estimator $\hat{f}$ and symmetric property f, there exists a symmetric set $\mathrm{P}' \subseteq \Delta$ such that for some $p \in \mathrm{P}'$,

$$r_n(p, \hat{f}\,) \ge 2r_{n_p}\big(\mathrm{P}'\big)\,.$$

We provide some brief comments here and more in Section 5.3. First, the above theorem holds for a polynomial-time PML approximation [ACSS20], and for any symmetric property, while nearly all previous works require the property to possess certain forms and be smooth. In particular, the algorithm in [ACSS20] achieves the best-known guarantees for approximating PML, requires no additional assumptions on the distribution/property's structure, and works universally on all symmetric properties and adaptively on all profiles (hence distributions). Second, the result holds for any symmetric distribution set $P \subseteq \Delta$, which covers numerous domains of interest that appeared in the literature, such as the widely studied Δ, and its subset $\Delta_{1/|\mathcal{X}|}$ for the study of support size estimation, where each distribution's positive probabilities are at least $1/|\mathcal{X}|$. Third, the result trivially implies a weaker version in [ADOS17] where $\mathcal{H}_n(p)$ is replaced by $\sqrt{n}$, which, as we show in Section 5.2.5, can be significantly larger.

5.2.4 Optimal Compression of Profiles

None of the scientific applications in Section 5.1 is possible without first storing the sample profile.

Hence, we focus on the task of lossless profile compression in this section. Besides the theoretical fundamentality and numerous applications, the task is essential as storing a sample's profile, compared with storing the entire sample sequence, often takes much less space. Specifically, Shannon entropy is the measure of limit of lossless compression, which, for sample $X^n \sim p \in \Delta$, is $n\mathrm{H}(p)$, and for the sample's profile, is $\mathcal{H}_n(p)$. In particular, the sample entropy grows as $\Omega(n)$ whenever p has an entropy of at least one, while the profile entropy is at most $3\sqrt{n}$ by our argument in Section 5.2.2.

While the n-to-$\sqrt{n}$ improvement is already significant, the compression schemes we propose under the standard block and sequential settings surely take profile compression to the next level. Specifically, for every distribution p and sample size n, both schemes essentially compress the sample profile $\varphi(X^n)$ to its entropy $\mathcal{H}_n(p)$, the information-theoretic limit, in expectation. In other words, our algorithms are *instance-by-instance optimal* and essentially *unimprovable*. Furthermore, we achieve this instance optimality with *near-optimal time complexity* – both algorithms have a running time near-linear in the sample size n.

Block compression We propose an intuitive and easy-to-implement block compression algorithm.

Recall that the profile of a sequence x^n is the multiset $\varphi(x^n)$ of multiplicities associated with symbols in x^n. The ordering of elements in a multiset is not informative. Hence equivalently, we can compress $\varphi(x^n)$ into the set $\mathcal{C}(\varphi(x^n))$ of corresponding multiplicity-prevalence pairs, i.e.,

$$\mathcal{C}(\varphi(x^n)) := \{(\mu, \varphi_\mu(x^n)) : \mu \in \varphi(x^n)\}.$$

The number of pairs in $\mathcal{C}(\varphi(x^n))$ is equal to the profile dimension $\mathcal{D}(\varphi(x^n))$. Besides, both

prevalence and its multiplicity are integers in $[0, n]$, and storing the pair takes $2\log n$ nats. Hence, it takes at most $2(\log n) \cdot \mathcal{D}(\varphi(x^n))$ nats to store the compressed profile. By Theorem 1, for any distribution $p \in \Delta$ and sample $X^n \sim p$,

$$\mathbb{E}[2(\log n) \cdot \mathcal{D}(X^n)] \simeq \mathcal{H}_n(p).$$

We have shown that storing a profile φ as $\mathcal{C}(\varphi)$ is a near-optimal block compression scheme.

Sequential compression For any sequence x^n, the setting for sequential profile compression is that at time step $t \in [n]$, the compression algorithm knows only $\varphi(x^t)$ and sequentially encodes the new information. This process is equivalent to providing the algorithm $\mu_{x_t}(x^{t-1})$ at time step t.

Suppress x, x^t in the expressions for the ease of illustration. For efficient compression, we sequentially encode the profile φ into a *self-balancing binary search tree* $\mathcal{T}$, with each node storing a multiplicity-prevalence pair (μ, φ_μ) and μ being the search key. We present the compression scheme as Algorithm 1, and establish the following guarantee.

Theorem 5. *Algorithm 1 runs for exactly n iterations, with an $O(\log n)$ per-iteration time complexity. For an i.i.d.sample $X^n \sim p$, the expected space complexity is $\tilde{\Theta}(\mathcal{H}_n(p))$. On the other hand, any algorithm that compresses the profile losslessly has an expected space complexity of at least $\mathcal{H}_n(p)$.*

5.2.5 Optimal Characterization for Structured Families

In this section, we characterize the profile entropy of several important structured distribution families, including log-concave, power-law, histogram, and their mixtures. All the matching lower bounds are entirely new, and all the upper bounds, with the exception of that in Theorem 8, are much stronger than those induced by the prior work [HO19b] via Theorem 1. For interested readers, see Appendix 5.D for a detailed comparison.

Algorithm 1 Sequential Profile Compression

input sequence $(\mu_{x_t}(x^{t-1}))_{t=1}^{n}$, tree $\mathcal{T} = \varnothing$
output tree $\mathcal{T}$ that encodes the input sequence
for t = 1 to n **do**
  if $\mu := \mu_{x_t}(x^{t-1}) \in \mathcal{T}$ **then**
    if $\mu+1 \in \mathcal{T}$ **then**
      $\varphi_{\mu+1} := \mathcal{T}(\mu+1) \leftarrow \mathcal{T}(\mu+1)+1$
    else
      add $(\mu+1, 1)$ to $\mathcal{T}$
    end if
    if $\varphi_\mu = 1$ **then** delete (μ, φ_μ) from $\mathcal{T}$
    else $\varphi_\mu := \mathcal{T}(\mu) \leftarrow \mathcal{T}(\mu) - 1$ **endif**
  else
    if $1 \notin \mathcal{T}$ **then** add $(1,1)$ to $\mathcal{T}$
    else $\mathcal{T}(1) \leftarrow \mathcal{T}(1)+1$ **endif**
  end if
end for

Log-concave The log-concave family encompasses a broad range of discrete distributions, such as Poisson, hyper-Poisson, Poisson binomial, binomial, negative binomial, and geometric, and hyper-geometric, with broad applications to statistics [SW14], computer science [LV07], economics [An97], and geometry [Sta89].

Formally, a distribution $p \in \Delta_{\mathbb{Z}}$ is *log-concave* if p has a contiguous support and $p_x^2 \geq p_{x-1} \cdot p_{x+1}$ for all $x \in \mathbb{Z}$. The next result bounds the profile entropy of this family, and is *tight* up to logarithmic factors. For simplicity, henceforth we write $a \wedge b$ for $\min\{a,b\}$ (and $\vee$ for max), and slightly abuse the notation and write $a \simeq b$ for $a+1 = \tilde{\Theta}(b+1)$, which does not change the nature of the results.

Theorem 6. *Let $\mathcal{L}_\sigma \subseteq \Delta_{\mathbb{Z}}$ denote the collection of log-concave distributions with variance σ^2. Then,*

$$\max_{p \in L_\sigma} \mathcal{H}_n(p) \simeq \sigma \wedge \frac{n}{\sigma}.$$

In particular, if we discretize a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma^2)$ by rounding it to the nearest integer, the distribution of the resulting variable achieves the maximum, up to logarithmic factors.

Moreover, such a discretization procedure preserves log-concavity for any continuous distribution over $\mathbb{R}$.

Power-law Power-law is a ubiquitous structure appearing in many situations of scientific interest, ranging from natural phenomena such as the initial mass function of stars [Kro01], species and genera [HQD+10], rainfall [MR93], population dynamics [Tay61], and brain surface electric potential [MSODN09], to human-made circumstances such as the word frequencies in a text [Baa02], income rankings [DY01], company sizes [Axt01], and internet topology [FFF99].

Formally, a discrete distribution $p \in \Delta_{\mathbb{Z}}$ is a *power-law with power* $\alpha \geq 0$ if p has a support of $[k] := \{1, \ldots, k\}$ for some $k \in \mathbb{Z}^+ \cup \{\infty\}$ and $p_x \propto x^{-\alpha}$ for all $x \in [k]$. Note that if $\alpha \in [0,1]$, the distribution is well-defined for only finite k. The next result fully characterizes the profile entropy of power-laws over all α, n, and k ranges, and significantly improves that in [HO19b].

Theorem 7. *Let* $p \in \Delta_{[k]}$ *be a power-law distribution with power* α. *Then,*

$$\mathcal{H}_n(p) \simeq \begin{cases} k & \text{if } \alpha > \frac{k^{1+\alpha}}{n} \vee 1 \text{ or } 1 \geq \alpha > \frac{k^2}{n}, \\ n^{\frac{1}{\alpha+1}} & \text{if } \frac{k^{1+\alpha}}{n} \geq \alpha > 1, \\ \left(\frac{n}{k^{1-\alpha}}\right)^{\frac{1}{1+\alpha}} & \text{if } \frac{k^2}{n} \wedge 1 \geq \alpha > \frac{k^{1-\alpha}}{n}, \\ \frac{n}{k^{1-\alpha}} - \frac{n}{k} & \text{if } \frac{k^{1-\alpha}}{n} \wedge 1 \geq \alpha \text{ and } \alpha \geq 2\log_k\left(7\sqrt{\frac{k}{n}}+1\right), \\ k \wedge \sqrt{\frac{n}{k^{1-\alpha}}} & \text{if } \frac{k^{1-\alpha}}{n} \wedge 1 \geq \alpha \text{ and } 2\log_k\left(7\sqrt{\frac{k}{n}}+1\right) > \alpha. \end{cases}$$

In particular, as $\alpha \to 0$, *the bound degenerates to* $k \wedge \sqrt{\frac{n}{k}}$, *which is at most* $n^{\frac{1}{3}}$.

Since a power-law sample profile is completely specified by α, k, and n, the above theorem directly applies to model parameter estimation. Specifically, we first compute $\mathcal{D}_n \sim p$, which is a simple function of the symbol counts. By Theorem 1, we can then use it to approximate $\mathcal{H}_n(p)$. Finally, we utilize the characterization theorem and find the parameter relations (testing might be necessary).

Histogram While histogram is among the most widely studied representations, histogram distributions' importance also rises with the rapid growth of data sizes in modern scientific applications. For example, *subsampling*, a generic strategy to handle large datasets, naturally induces a histogram distribution over different categories of the data. This induced distribution often summarizes vital data statistics, leveraging which yields efficient and flexible inference procedures.

Formally, a discrete distribution $p \in \Delta_{\mathbb{Z}}$ is a *t-histogram* if we can partition its support into at most t pieces such that p takes the same probability value over each piece. The theorem below provides near-optimal bounds on the profile entropy of the t-histogram distributions.

Theorem 8. *Denote by $I_t \subseteq \Delta_{\mathbb{Z}}$ the collection of t-histogram distributions. Then,*

$$\max_{p \in I_t} \mathcal{H}_n(p) \simeq (nt^2)^{\frac{1}{3}} \wedge \sqrt{n}.$$

In practical settings, the value of t is often poly-logarithmic in n, and the bound reduces to $\tilde{O}(n^{1/3})$. For the particular case of $t = 1$, distribution p is uniform over some unknown contiguous support. This result overlaps with Theorem 7 with $\alpha = 0$, yielding the following bound.

Corollary 3. *For any uniform distribution p with support size k, we have $\mathcal{H}_n(p) \simeq k \wedge \sqrt{\frac{n}{k}}$.*

5.3 Applications and Extensions

Robust learning The profile of any sequence is invariant to domain-symbol permutations. Since entropy is a symmetric property, the profile entropy of an i.i.d.sample is also permutation invariant. Consequently, a result in this paper that holds for a distribution will also hold for *any distributions possessing the same probability multiset*. For numerous practical applications, this *robustness to symbol permutation* is a desirable and novel notion of robustness that particularly resides in discrete domains, as samples often come as categorical data, while the alphabet ordering for the underlying distribution to exhibit certain structure is frequently unknown [HO19b].

For example, the sample may consist of different fruits, not integers. But suppose there is a hidden mapping from the fruit domain to integers that makes the distribution log-concave over $\mathbb{Z}$. Then, all our results such as Theorem 2, 4, 5, and 6 are in effect. For another example, in natural language processing, we observe words and punctuation marks. Even we know that observations come from a power-law distribution [Mit04], it is often unclear how to order the alphabet to realize such a condition. The robustness of our approach again enables us to achieve a variety of learning objectives, such as understanding the relation between different model parameters (Theorem 7).

Mixture models The results in Section 5.2.5 provide optimal characterization for simple structured families. A standard extension to incorporate more complex structures in the model is spanning a distribution family by including (weighted) mixtures. A typical example is the Gaussian mixture model, which is among the most widely studied probabilistic models.

In the supplementary material, we present such results for all three families in Section 5.2.5, and for mixtures of discretized high-dimensional Gaussians. In fact, we obtain a simple and intuitive profile-entropy characterization for all distributions. Partition the unit interval into a sequence of ranges, $I_j := \left((j-1)^2\frac{\log n}{n},\ j^2\frac{\log n}{n}\right], 1 \le j \le \sqrt{\frac{n}{\log n}}$, and for any distribution p, denote by p_{I_j} the number of probabilities in I_j. Then,

Lemma 1. *For any $n \in \mathbb{Z}^+$ and $p \in \Delta$, we have $\mathcal{H}_n(p) \simeq \sum_{j\ge 1} \min\left\{p_{I_j}, j\cdot\log n\right\}$.*

Competitive property estimation Theorem 2 on PML holds for every distribution, any symmetric property, and distribution collection, such as a finite-dimensional simplex, regardless of other parameters such as the alphabet size. To the best of our knowledge, this is one of the most general results in the field. Below we provide a basic example for its applications.

For an arbitrary $\beta > 0$, let f be the order-β Rényi entropy, and P be the set of distributions whose probability multisets correspond to power-laws with power $\alpha \ge 3$. The minimax error rate $r_m(\mathrm{P})$ is unknown for this problem as recent works (e.g., [AOST16]) mainly focused on the standard simplexes. On the other hand, Theorem 4, together with Theorem 7, shows that the

n-sample PML plug-in estimator essentially performs as well as the best $n^{3/4}$-sample estimator. Note that while the guarantee of PML uniformly holds for all β, the best estimator can optimize its performance for every β. Following the same rationale, we can derive such nontrivial competitive estimation results for numerous properties and distribution families without having to analyze them in detail.

Adaptive testing and classification Profile entropy also directly connects to adaptive testing and classification. Such a connection arises from computing the *profile probability* [ADJ$^+$11, ADJ$^+$12b], the probability of observing the sample's profile under the same sampling process.

Specifically, the first paper designs an algorithm that distinguishes two unknown distributions using near-optimal sample sizes whenever the optimal algorithm has an exponentially small error probability. In addition, the algorithm is simply a ratio test between the probabilities of two profiles. Given sample $X^n \sim p$ over a finite domain, we can compute its profile probability in $\exp(\tilde{\Theta}(\mathcal{H}_n(p)))$ operations. For example, if the underlying distribution is a 4-histogram, then by Theorem 8, the running time exponent is of order $n^{1/3}$. The result follows by the equivalence of the problem and computing the permanent of a rank-$\mathcal{D}_n$ matrix [Bar96, Von12, Von14, Bar16].

Method of types We connect our approach to *the method of types*, an important technical tool in Shannon theory and many other fields [CK11, Wol12]. In the notation of this paper, the *type* of a sequence x^n over some finite domain $\mathcal{X}$ is the ordered list of multiplicities $\mu_y(x^n)$, which associates symbol y with its number of appearances in x^n. For this multiplicity list, the method of types associates each $\mu_y(x^n)$ with the number of symbols having this multiplicity, which is precisely $\varphi_{\mu_y(x^n)}(x^n)$. Hence, the profile of a sequence is *the type of its type*.

Given the above arguments, understanding the deep connection between profile-based algorithms and the method of types is a meaningful future research direction to explore.

5.4 Conclusion and Broader Impact

Classical information theory states that an i.i.d.sample contains $H(X^n \sim p) = nH(p)$ information, which provides little insight for statistical applications. We present a different view by decomposing the sample information into three parts: the labeling of the profile elements, ordering of them, and profile entropy. With no bias towards any symbols, the *profile entropy* rises as a fundamental measure unifying the concepts of estimation, inference, and compression. We believe this view could help researchers in information theory, statistical learning theory, and computer science communities better understand the information composition of i.i.d.samples over discrete domains.

The results established in this work are general and fundamental, and have numerous applications in privacy, economics, data storage, supervised learning, etc. A potential downside is that the theoretical guarantees of the associated algorithms rely on the assumption correctness, e.g., the domain should be discrete and the sampling process should be i.i.d.. In other words, it will be better if users can confirm these assumptions by prior knowledge, experiences, or statistical testing procedures. Taking a different perspective, we think a potential research direction following this work is to extend these results to Markovian models, making them more robust to model misspecification.

5.5 Acknowledgements

Chapter 6, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Yi Hao and Alon Orlitsky, 2020. The dissertation author was the primary investigator and author of this paper.

Appendix organization In the appendix, we order the results and proofs according to their logical priority. In other words, the proof of a theorem or lemma mainly relies on preceding results. For the ease of reference, the numbering of the theorems is consistent with that in the

main paper.

5.A Entropy and Dimension of Sample Profiles

Consider an arbitrary sequence x^n over a finite or countably infinite alphabet $\mathcal{X}$. The *multiplicity* $\mu_y(x^n)$ of a symbol $y \in \mathcal{X}$ is the number of times y appears in x^n. The *prevalence* of an integer μ is the number $\varphi_\mu(x^n)$ of symbols in x^n with multiplicity μ. The *profile* of x^n is the multiset $\varphi(x^n)$ of multiplicities of the symbols in x^n. We refer to it as a profile of *length* n.

The number $\mathcal{D}(S)$ of distinct elements in a multiset S is its *dimension*. For convenience, we also write $\mathcal{D}(x^n)$ for profile dimension. The dimension of a length-n profile over $\mathcal{X}$ is at most $\min\{\sqrt{2n}, |\mathcal{X}|\}$.

Let Δ be the collection of all discrete distributions, and Δ be the collection of those over $\mathcal{X}$. Draw a size-n sample X^n from an arbitrary distribution in $p \in \Delta$. Then, the profile Φ^n of X^n is a random multiset whose distribution depends on only p and n. We therefore write $\Phi^n \sim p$, and call $\mathcal{H}_n(p) := H_n(p)$ the *profile entropy* with respect to (p,n). Analogously, we call $\mathcal{D}_n := \mathcal{D}_n$, the *profile dimension* associated with (p,n), and write $\mathcal{D}_n \sim p$.

Consider an arbitrary sequence x^n over a finite or countably infinite alphabet $\mathcal{X}$. The *multiplicity* $\mu_y(x^n)$ of a symbol $y \in \mathcal{X}$ is the frequency of y in x^n. The *prevalence* of an integer μ is the number $\varphi_\mu(x^n)$ of symbols in x^n with multiplicity μ. The *profile* of x^n is the multiset $\varphi(x^n)$ of multiplicities of the symbols in x^n, which we describe as a profile of *length* n.

5.A.1 Concentration of Profile Dimension

First we express the dimension of a sample profile in terms of the symbol multiplicities. Denote by $\bigvee$ the logical OR operator. For any distribution p and $X^n \sim p$,

$$\mathcal{D}_n = \sum_{\mu=1}^{n} \bigvee_{x \in \mathcal{X}} \mathbb{1}_{\mu_x(X^n)=\mu}.$$

The statistical dependency landscape of terms in the summation is rather complex, since $\mu_x(X^n)$ and $\mu_y(X^n)$ are dependent for every (x,y) pair due to the fixed sample size; and so are $\mathbb{1}_{\mu_x(X^n)=\mu_1}$ and $\mathbb{1}_{\mu_x(X^n)=\mu_2}$ for every pair of distinct μ_1 and μ_2. To simplify the derivations, we relate this quantity to its variant under the *Poisson sampling scheme*, i.e., making the sample size an independent $N \sim \mathrm{Poi}(n)$. Specifically, define

$$\tilde{\mathcal{D}}_N := \tilde{\mathcal{D}}(X^N) := \sum_{U=1}^{n} \bigvee_{x\in\mathcal{X}} \mathbb{1}_{\mu_x(X^N)=U}.$$

Note that this is not the same as $\mathcal{D}_N$ since the summation index goes up only to n.

Denote the expected value of $\tilde{\mathcal{D}}_N$ by $E_n(p)$, which will frequently appear in the rest discussions. Our first result shows that the original $\mathcal{D}_n$ satisfies a Chernoff-Hoeffding type bound centered at $E_n(p)$.

Theorem 9. *Under the above conditions and for any $n \in \mathbb{Z}^+$, $p \in \Delta$, and $\gamma > 0$,*

$$\Pr\left(\frac{\mathcal{D}_n}{1+\gamma} \geq E_n(p)\right) \leq 3\sqrt{n}e^{-\min\{\gamma^2,\gamma\}E_n(p)/3},$$

and for any $\gamma \in (0,1)$,

$$\Pr\left(\frac{\mathcal{D}_n}{1-\gamma} \leq E_n(p)\right) \leq 3\sqrt{n}e^{-\gamma^2 E_n(p)/2}.$$

Proof. A nice attribute of Poisson sampling is that all the multiplicities $\mu_y(X^n)$ are independent of each other. We will first consider $\mathcal{D}_N$ and relate it to the fixed-sample-size version later.

For simplicity and clarity, we suppress X^n in $\mu_y(X^n)$ and write ν_y instead of μ_y when the multiplicity is obtained through Poisson sampling. For any $i \in [n]$, denote $G_i(\{\nu_x\}_x) := \bigvee_{x\in\mathcal{X}} \mathbb{1}_{\nu_x=i}$. As mentioned previously, instead of analyzing $\mathcal{D}_N$, we consider

$$\tilde{\mathcal{D}}_N = \sum_{i=1}^{n} \bigvee_{x\in\mathcal{X}} \mathbb{1}_{\nu_x=i} = \sum_{i=1}^{n} G_i(\{\nu_x\}_x).$$

Note that for any disjoint $I, J \subseteq [n]$, the functions $\sum_{i\in I} G_i(\{\nu_x\}_x)$ and $\sum_{j\in J} G_j(\{\nu_x\}_x)$ are discordant monotone by each argument, namely, when we increase the value of each ν_x, the increase in the value of one function implies the non-increase of the other. Then, by the results in [Leh66], the values of the two functions, when viewed as random variables, are negatively associated.

Next we show that quantity $\tilde{\mathcal{D}}_N$ satisfies a Chernoff-type bound.

Let γ be an arbitrary positive number. Note that G_i is a Bernoulli random variable with parameter $q_i := \mathbb{E}\left[G_i(\{\nu_x\}_x)\right]$. Then for the expected value of $\tilde{\mathcal{D}}_N$, we have

$$E_n(p) := \mathbb{E}\left[\tilde{\mathcal{D}}_N\right] = \mathbb{E}\left[\sum_{i=1}^{n} G_i(\{\nu_x\}_x)\right] = \sum_i q_i.$$

For simplicity, temporally write $Y := \tilde{\mathcal{D}}_N$ and $\mu := E_n(p)$. Then, by Markov's inequality and the monotonicity of function e^{ty} over $t > 0$,

$$\Pr\left(Y \geq (1+\gamma)\mu\right) = \Pr\left(e^{tY} \geq e^{t(1+\gamma)\mu}\right) \leq \frac{\mathbb{E}[e^{tY}]}{e^{t(1+\gamma)\mu}}.$$

It suffices to bound $\mathbb{E}[e^{tY}]$ by a function of other parameters.

$$\begin{aligned}
\mathbb{E}[e^{tY}] &\overset{(a)}{=} \mathbb{E}\left[\exp\left(t\left(\sum_{i=1}^{n} G_i(\{M_x\}_x)\right)\right)\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\exp\left(tG_1(\{M_x\}_x)\right)\cdot\exp\left(t\left(\sum_{i=2}^{n} G_i(\{M_x\}_x)\right)\right)\right] \\
&\overset{(c)}{\leq} \mathbb{E}\left[\exp\left(tG_1(\{M_x\}_x)\right)\right]\cdot\mathbb{E}\left[\exp\left(t\left(\sum_{i=2}^{n} G_i(\{M_x\}_x)\right)\right)\right] \\
&\overset{(d)}{\leq} \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(tG_i(\{M_x\}_x)\right)\right] \overset{(e)}{=} \prod_{i=1}^{n}\left(1+q_i(e^t-1)\right) \\
&\overset{(f)}{\leq} \prod_{i=1}^{n}\left(\exp\left(q_i(e^t-1)\right)\right) \overset{(g)}{=} \exp\left(\sum_{i=1}^{n} q_i(e^t-1)\right) \\
&\overset{(h)}{=} \exp\left((e^t-1)\mu\right),
\end{aligned}$$

where (a) follows by the definition of Y; (b) follows by $e^{a+b} = e^a \cdot e^b$; (c) follows by the fact that G_1 is negatively associated with $\sum_{i=2}^{n} G_i$; (d) follows by an induction argument via negative association; (e) follows by the fact that G_i is a Bernoulli random variable with mean q_i; (f) follows by the inequality $1 + x \leq e^x, \forall x \geq 0$; (g) follows by $e^a \cdot e^b = e^{a+b}$; and (h) follows by $\mu = \sum_i q_i$.

Applying standard simplifications, we obtain

$$\Pr\left(Y \geq (1+\gamma)\mu\right) \leq e^{-\min\{\gamma^2,\gamma\}\mu/3}, \ \forall \gamma > 0,$$

and

$$\Pr\left(Y \leq (1-\gamma)\mu\right) \leq e^{-\gamma^2\mu/2}, \ \forall \gamma \in (0,1).$$

The proof will be complete upon noting that: 1) the probability that $N = n$ is at least $1/(3\sqrt{n})$; 2) conditioning on $N = n$ transforms the sampling model to that with a fixed sample size n. □

As a corollary, the value of $\mathcal{D}_n$ is often close to $E_n(p)$.

Corollary 4. *Under the same conditions as above and for any $n \in \mathbb{Z}^+$, $p \in \Delta$, with probability at least $1 - 6/\sqrt{n}$,*

$$\frac{1}{2}E_n(p) - 4\log n \leq \mathcal{D}_n \leq 2E_n(p) + 3\log n.$$

Proof. To establish the lower bound, note that if $E_n(p) \geq 3\log n$, setting $\gamma = 1$ in Theorem 9 yields

$$\Pr\left(\mathcal{D}_n \geq 2E_n(p) + 3\log n\right) \leq \Pr\left(\mathcal{D}_n \geq 2E_n(p)\right) \leq 3\sqrt{n}e^{-E_n(p)/3} \leq \frac{3}{\sqrt{n}},$$

else if $E_n(p) < 3\log n$, setting $\gamma = (3\log n)/E_n(p)$ yields

$$\Pr\left(\mathcal{D}_n \geq 2E_n(p) + 3\log n\right) \leq \Pr\left(\mathcal{D}_n \geq E_n(p) + 3\log n\right) \leq 3\sqrt{n}e^{-(3\log n)/3} = \frac{3}{\sqrt{n}}.$$

As for the upper bound, if $E_n(p) \geq 8 \log n$,

$$\Pr\left(\mathcal{D}_n + 4\log n \leq \left(1 - \frac{1}{2}\right) E_n(p)\right) \leq \Pr\left(\mathcal{D}_n \leq \left(1 - \frac{1}{2}\right) E_n(p)\right) \leq 3\sqrt{n} e^{-\mu/8} \leq \frac{3}{\sqrt{n}},$$

and for any $E_n(p) < 8 \log n$,

$$\Pr\left(\mathcal{D}_n + 4\log n \leq \left(1 - \frac{1}{2}\right) E_n(p)\right) \leq \Pr\left(\mathcal{D}_n < 0\right) = 0 \leq \frac{3}{\sqrt{n}}.$$

Combining these tail bounds through the union bound completes the proof. □

In addition to the above, we establish an Efron-Stein type inequality.

Theorem 10. *For any distribution p and $\mathcal{D}_n \sim p$,*

$$\mathrm{Var}(\mathcal{D}_n) \leq \mathbb{E}[\mathcal{D}_n].$$

Proof. First, note that for any $j, t \in [n]$ and $j \neq t$,

$$\begin{aligned}
C_{j,t} &:= \mathrm{Cov}\left(\mathbb{1}_{\varphi_j(X^n)>0}, \mathbb{1}_{\varphi_t(X^n)>0}\right) \\
&= \Pr\left(\varphi_j(X^n), \varphi_t(X^n) > 0\right) - \Pr\left(\varphi_j(X^n) > 0\right) \cdot \Pr\left(\varphi_t(X^n) > 0\right) \\
&= \left(\Pr\left(\varphi_j(X^n) > 0 | \varphi_t(X^n) > 0\right) - \Pr\left(\varphi_j(X^n) > 0\right)\right) \cdot \Pr\left(\varphi_t(X^n) > 0\right) \\
&= \left(\Pr\left(\varphi_j(X^n) > 0 | \varphi_t(X^n) > 0\right) - \Pr\left(\varphi_j(X^n) > 0 | \varphi_t(X^n) = 0\right)\right) \\
&\quad \times \Pr\left(\varphi_t(X^n) = 0\right) \cdot \Pr\left(\varphi_t(X^n) > 0\right) \\
&\leq 0
\end{aligned}$$

Therefore, the variance of the profile dimension $\mathcal{D}_n$ satisfies

$$\begin{aligned}
\mathrm{Var}\left(\mathcal{D}_n\right) &= \mathrm{Var}\left(\sum_{i=1}^{n} \mathbb{1}_{\varphi_i(X^n)>0}\right) \\
&\leq \sum_{i=1} \mathrm{Var}\left(\mathbb{1}_{\varphi_i(X^n)>0}\right) + \sum_{j\neq t} \mathrm{Cov}\left(\mathbb{1}_{\varphi_j(X^n)>0}, \mathbb{1}_{\varphi_t(X^n)>0}\right) \\
&\leq \sum_{i=1} \mathbb{E}\left[\mathbb{1}_{\varphi_i(X^n)>0}\right] + \sum_{j\neq t} C_{j,t} \\
&\leq \sum_{i=1} \mathbb{E}\left[\mathbb{1}_{\varphi_i(X^n)>0}\right] \\
&= \mathbb{E}\left[\mathcal{D}_n\right]. \qquad \square
\end{aligned}$$

5.A.2 Theorem 1: Dimension-Entropy Equivalence

The following theorem shows that for every distribution and sampling parameter n, the induced profile entropy and dimension are of the same order, with high probability.

Theorem 1 (Entropy-dimension equivalence)**.** *For any distribution $p \in \Delta$ and $\mathcal{D}_n \sim p$,*

$$\Pr(\mathcal{D}_n \simeq \mathcal{H}_n(p)) \geq 1 - \frac{1}{\sqrt{n}}.$$

5.A.3 Proof of Theorem 1

Proof outline We decompose the proof of the theorem into three steps.

First, we show $\mathcal{H}_n(p) \lesssim \mathcal{D}_n$ with high probability, which is a consequence of Theorem 9 (which shows that $\mathcal{D}_n$ highly concentrates around its expectation) and Shannon's source coding theorem. Second, we introduce a simple quantity $\mathcal{H}_n^s(p)$ that approximates the expectation of $\mathcal{D}_n$ to within logarithmic factors of n. Finally, leveraging this approximation guarantee, we establish the other direction of the theorem. This step is more involved due to the aforementioned complications.

Step 1: Bounding Profile Entropy by Its Dimension

By the tail bounds (Theorem 9) and trivial lower bound of 1 on the profile dimension, with probability at least $1 - 1/\sqrt{n}$, the expectation of $\mathcal{D}_n$ satisfies

$$\mathbb{E}[\mathcal{D}_n] \lesssim \mathcal{D}_n.$$

By our result on block profile compression (Section 5.2.4), storing profile $\Phi^n \sim p$ losslessly takes

$$O(\log n) \cdot \mathbb{E}[\mathcal{D}_n] + O\left(\frac{1}{\sqrt{n}}\right) \cdot \log \mathbb{P}(n) \lesssim \mathbb{E}[\mathcal{D}_n]$$

nats space in expectation. By Shannon's source coding theorem, the expected space to losslessly storing a random variable is at least its entropy. Hence, with probability at least $1 - O(1/\sqrt{n})$,

$$\mathcal{H}_n(p) \lesssim \mathbb{E}[\mathcal{D}_n] \lesssim \mathcal{D}_n.$$

Applying $\mathcal{D}_n \geq 1$ completes the proof.

Step 2: Simple Approximation Formula for Profile Dimension

Next, we show that $\mathcal{H}_n(p) \gtrsim \mathcal{D}_n$, with high probability. Note that $\mathcal{D}_n \sim p$ is often close to $E_n(p)$, the expectation of its Poissonized version $\tilde{\mathcal{D}}_N$, with an exponentially small deviation probability. Hence, to approximate $\mathcal{D}_n$, it suffices to accurately compute $E_n(p)$.

By independence and the linearity of expectations,

$$E_n(p) = \mathbb{E}[\tilde{\mathcal{D}}_N] = \sum_{i=1}^{n} \left(1 - \prod_{x \in \mathcal{X}} \left(1 - e^{-np_x} \frac{(np_x)^i}{i!}\right)\right).$$

The expression is exact but does not relate to p in a simple manner. For an intuitive approximation,

we partition the unit interval into a sequence of ranges,

$$I_j := \left((j-1)^2 \frac{\log n}{n},\ j^2 \frac{\log n}{n}\right], 1 \le j \le \sqrt{\frac{n}{\log n}},$$

denote by p_{I_j} the number of probabilities p_x belonging to I_j, and relate $E_n(p)$ to an induced shape-reflecting quantity,

$$\mathcal{H}_n^s(p) := \sum_{j \ge 1} \min\left\{p_{I_j}, j \cdot \log n\right\},$$

the sum of the effective number of probabilities lying within each range [HO19b]. To compute $\mathcal{H}_n^s(p)$, we simply count the number of probabilities in each I_j. Our main result shows that $\mathcal{H}_n^s(p)$ well approximates $E_n(p)$ over the entire Δ, up to logarithmic factors of n.

Theorem 11. *For any $n \in \mathbb{Z}^+$ and $p \in \Delta$,*

$$\frac{1}{\sqrt{\log n}} \cdot \Omega(\mathcal{H}_n^s(p)) \le E_n(p) \le O(\mathcal{H}_n^s(p)).$$

Proof. The fact that $O(H_n^S(p))$ upperly bounds $\mathbb{E}[\tilde{\mathcal{D}}_N]$ simply follows by the concentration of Poisson variables, and is established in [HO19b]. Below we show that the quantity also serves as a lower bound. By construction, for any given sampling parameter n, index j, and symbol x with probability $p_x \in I_j$, the corresponding symbol multiplicity $\mu_x \sim \text{Poi}(np_x)$. Hence, we can express the expectation of $\tilde{\mathcal{D}}_N$ as

$$\begin{aligned}
\mathbb{E}\left[\tilde{\mathcal{D}}_N\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \bigvee_x \mathbb{1}_{\mu_x = i}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[1 - \bigwedge_x \mathbb{1}_{\mu_x \neq i}\right] \\
&= \sum_{i=1}^{n} \left(1 - \prod_x \mathbb{E}\left[\mathbb{1}_{\mu_x \neq i}\right]\right) \\
&= \sum_{i=1}^{n} \left(1 - \prod_x \left(1 - e^{-np_x} \frac{(np_x)^i}{i!}\right)\right).
\end{aligned}$$

This proves the aforementioned formula. Then, for every sufficiently large index j and $i \in S_j := [(j-1)^2, j^2]\log n$, define a sequence of intervals,

$$I_j^i := \frac{i}{n} + [-j, j]\,\frac{\sqrt{\log n}}{n}.$$

Then for any $i \in S_j$ and $p_x \in I_j^i \cap I_j$, the corresponding Poisson probability satisfies

$$\begin{aligned}
e^{-np_x}\frac{(np_x)^i}{i!} &= e^{-i}\frac{i^i}{i!}\cdot\left(e^{i-np_x}\cdot\frac{(np_x)^i}{i^i}\right)\\
&= e^{-i}\frac{i^i}{i!}\cdot\left(e^{-(np_x-i)}\cdot\left(1+\frac{np_x-i}{i}\right)^i\right)\\
&= e^{-i}\frac{i^i}{i!}\cdot\exp\left(-(np_x-i)+i\cdot\log\left(1+\frac{np_x-i}{i}\right)\right)\\
&\geq \frac{1}{3\sqrt{i}}\cdot\exp\left(-\frac{2i}{3}\cdot\left(\frac{np_x-i}{i}\right)^2\right)\\
&\geq \frac{1}{9\sqrt{i}} \geq \frac{1}{9j\sqrt{\log n}}.
\end{aligned}$$

Now we analyze the contribution of indices $i \in S_j$ to the expected value of $\tilde{\mathcal{D}}_N$. For clarity, we divide our analysis into two cases: $p_{I_j} \geq j\log n$ and $p_{I_j} < j\log n$.

Consider the collection $\mathcal{P}_j$ of probabilities $p_x \in I_j$, and the collection I_j of intervals $I_j^i, i \in S_j$. By construction, each probability in $\mathcal{P}_j$ is contained in at least $j\sqrt{\log n}$ many intervals in I_j. Hence the total number of probabilities (repeatedly counted) included in I_j is at least $p_{I_j}\cdot j\sqrt{\log n}$. Note that the number of intervals in I_j is less than $2j\log n$. We claim that there exists one (or more) interval $I_j^{i'} \in I_j$ containing at least $p_{I_j}/(2\sqrt{\log n})$ probabilities. By construction, there are at least $j\sqrt{\log n}/2$ neighboring intervals of $I_j^{i'}$ that contain at least $p_{I_j}/(4\sqrt{\log n})$ probabilities.

The contribution of these these intervals to the expected value of $\tilde{\mathcal{D}}_N$ is at least $j\sqrt{\log n}/2$ times

$$\begin{aligned}
1-\left(1-\frac{1}{9j\sqrt{\log n}}\right)^{\frac{p_{I_j}}{4\sqrt{\log n}}} &\geq 1-\exp\left(\frac{p_{I_j}}{4\sqrt{\log n}}\log\left(1-\frac{1}{9j\sqrt{\log n}}\right)\right)\\
&\geq 1-\exp\left(-\frac{p_{I_j}}{40j\log n}\right)\\
&\geq \Theta\left(\frac{p_{I_j}}{j\log n}\right),
\end{aligned}$$

where the last step holds if $p_{I_j} \leq j\log n$. This yields a lower bound of $\Theta(p_{I_j}/\sqrt{\log n})$.

It remains to consider the $p_{I_j} > j\log n$ case. Again, the total number of probabilities included in I_j is at least $p_{I_j} \cdot j\sqrt{\log n}$. Furthermore, each interval I_j^i contains at most p_{I_j} probabilities and there are less than $2j\log n$ intervals. Therefore, the number of intervals that contain at least $j\sqrt{\log n}/4$ probabilities is at least $j\sqrt{\log n}/2$. Otherwise, the number of probabilities included in I_j is less than

$$\frac{j\sqrt{\log n}}{4}\cdot 2j\log n + p_{I_j}\cdot\frac{j\sqrt{\log n}}{2} \leq p_{I_j}\cdot j\sqrt{\log n},$$

which leads to a contradiction. Analogously, the contribution of these these intervals to the expected value of $\tilde{\mathcal{D}}_N$ is at least $j\sqrt{\log n}/2$ times

$$\begin{aligned}
1-\left(1-\frac{1}{9j\sqrt{\log n}}\right)^{\frac{j\sqrt{\log n}}{4}} &\geq 1-\exp\left(\frac{j\sqrt{\log n}}{4}\log\left(1-\frac{1}{9j\sqrt{\log n}}\right)\right)\\
&\geq 1-\exp\left(-\frac{1}{40}\right)\\
&= \Theta(1),
\end{aligned}$$

which yields a lower bound of $\Theta(j\sqrt{\log n})$ on the expected value of $\tilde{\mathcal{D}}_N$.

Consolidating the previous results shows that

$$\mathbb{E}\left[\tilde{\mathcal{D}}_N\right] \geq \frac{1}{\sqrt{\log n}} \cdot \Omega\Big(\sum_{j\geq 1} \min\left\{p_{I_j}, j\cdot \log n\right\}\Big). \qquad \square$$

Step 3: Bounding Profile Dimension by Its Entropy

Next, we establish that for any distribution $p \in \Delta$, $\Phi^n \sim p$, with probability at least $1 - 1/\sqrt{n}$,

$$\mathcal{H}_n(p) \gtrsim \mathcal{D}_n.$$

Let p be an arbitrary distribution in Δ. Recall that we partition the interval $(0,1]$ into a sequence of sub-intervals,

$$I_j := \left((j-1)^2\frac{\log n}{n}, j^2\frac{\log n}{n}\right], \quad 1 \leq j \leq \sqrt{\frac{n}{\log n}},$$

and denote by p_{I_j} the number of probabilities p_x in I_j.

Our current objective is to bound $H(\Phi^n \sim p)$ from below by a nontrivial multiple of $H_n^{\mathcal{S}}(p)$. For simplicity of derivations, we will adopt the standard Poisson sampling scheme and make the sample size an independent Poisson variable $N \sim \mathrm{Poi}(n)$. For notational simplicity, we suppress X^N in all the expressions and write the profile as $\varphi := \Phi^N$ by slightly abusing the notation.

Note that the profile can be equivalently expressed as a length-n vector

$$\varphi = (\varphi_1, \ldots, \varphi_n),$$

where φ_i denotes the number of symbols appearing exactly i times.

For a sufficiently large absolute constant c, decompose φ into c parts according to I_j such that the t-th part ($t = 1, \ldots, c$) consists of φ_i's satisfying $i \in nI_j$ with $j \equiv t \mod c$. Since by

definition,

$$H_n^{\mathcal{S}}(p) = \sum_{j \geq 1} \min\{p_{I_j}, j \cdot \log n\},$$

one of the c parts corresponds to a partial sum of at least $H_n^{\mathcal{S}}(p)/c$. Without loss of generality, we assume that it is the second part, i.e.,

$$\sum_{j \equiv 1 \mod c} \min\{p_{I_j}, j \cdot \log n\} \geq \frac{H_n^{\mathcal{S}}(p)}{c}.$$

Apply standard Poisson tail probability bounds. For example,

Lemma 2. *Let Y be a Poisson or binomial random variable with mean value λ. Then,*

$$\Pr(X \leq \lambda(1-\delta)) \leq \exp\left(-\frac{\delta^2\lambda}{2}\lambda\right), \ \ \forall \delta \in [0,1],$$

and

$$\Pr(X \geq \lambda(1+\delta)) \leq \exp\left(-\frac{\delta^2\lambda}{2+2\delta/3}\right), \ \ \forall \delta \geq 0.$$

For any $j \equiv 1 \mod c$ and with probability at least $1 - 1/n^4$, one can express the truncated profile $(\varphi_i)_{i \in nI_j}$ over I_j as a function of μ_x for x satisfying $np_x \in I_{j'}, j' \in (j - c/2, j + c/2)$.

Basically, this says that for every x, the number of its appearance is not too far away from the expected value. By the union bound, this is true for all $j \equiv 1 \mod c$ with probability at least $1 - 1/n^3$, as j can take only n possible values. Denote the last event by A.

To proceed, we recall the formula of [HR18] on the number $\mathbb{P}(n)$ of integer partitions of n, which happens to equal to the number of length-n profiles:

$$\log \mathbb{P}(n) = 2\pi\sqrt{\frac{n}{6}}(1 + o(1)).$$

Below, we will use a weaker version that works for any n:

$$\log \mathbb{P}(n) \leq \sqrt{3n}.$$

Then, conditioning on A, the truncated profiles $(\varphi_i)_{i \in nI_j}$ for $j \equiv 1 \mod c$ are independent. Since conditioning reduces entropy,

$$\begin{aligned}
H(\varphi) &\geq H((\varphi_i)_{i \in nI_j, j \equiv 1 \mod c}) \\
&\geq H((\varphi_i)_{i \in nI_j, j \equiv 1 \mod c} | \mathbb{1}_A) \\
&\geq H((\varphi_i)_{i \in nI_j, j \equiv 1 \mod c} | \mathbb{1}_A = 1) \cdot \Pr(A) \\
&= \sum_{j \equiv 1 \mod c} H((\varphi_i)_{i \in nI_j} | \mathbb{1}_A = 1) \cdot \Pr(A) \\
&= \sum_{j \equiv 1 \mod c} H((\varphi_i)_{i \in nI_j} | \mathbb{1}_A) - \sum_{j \equiv 1 \mod c} H((\varphi_i)_{i \in nI_j} | \mathbb{1}_A = 0) \cdot (1 - \Pr(A)) \\
&\geq \sum_{j \equiv 1 \mod c} (H((\varphi_i)_{i \in nI_j}) - H(\mathbb{1}_A)) - \frac{1}{n^3} \sum_{j \equiv 1 \mod c} H((\varphi_i)_{i \in nI_j} | \mathbb{1}_A = 0) \\
&\geq -nH(\mathbb{1}_A) + \sum_{j \equiv 1 \mod c} H((\varphi_i)_{i \in nI_j}) - \frac{1}{n^3} \cdot n \cdot \log(\exp(\Theta(\sqrt{n}))) \\
&= -O\left(\frac{1}{\sqrt{n}}\right) + \sum_{j \equiv 1 \mod c} H((\varphi_i)_{i \in nI_j}),
\end{aligned}$$

where the third last step follows by

$$H(X|Y) = H(X) - I(X,Y) = H(X) - H(Y) + H(Y|X) \geq H(X) - H(Y);$$

the second last follows by $H(X) \leq \log k$ for any X with a support size of k, and the fact that there are at most $\exp(3\sqrt{m})$ many profiles of length m, as we explained above; and the last step follows by the elementary inequality

$$H(\mathrm{Bern}(\theta)) \leq 2(\log 2)\sqrt{\theta(1-\theta)}, \ \forall \theta \in [0,1].$$

Our new objective is to bound $H((\varphi_i)_{i\in nI_j})$ from below. We will find a sub-interval I_j^s of I_j and bound $H((\varphi_i)_{i\in nI_j^s})$ in the rest of the section, since

$$H((\varphi_i)_{i\in nI_j}) \geq H((\varphi_i)_{i\in nI_j^s}).$$

For all $j \equiv 1 \mod c$, our lower bound is simply

$$H((\varphi_i)_{i\in nI_j^s}) \geq \Omega\left(\frac{1}{\sqrt{\log n}} \min\left\{p_{I_j}, j\cdot\log n\right\}\right),$$

which, together with $\sum_{j\equiv 1 \bmod c} \min\{p_{I_j}, j\cdot\log n\} \geq H_n^{\mathcal{S}}(p)/c$, implies that

$$H(\varphi) \geq -O\left(\frac{1}{\sqrt{n}}\right) + \sum_{j\equiv 1 \bmod c} H((\varphi_i)_{i\in nI_j}) \geq \Omega\left(\frac{1}{\sqrt{\log n}}\right)\cdot T_n.$$

Henceforth, we assume that j is sufficiently large and denote $L_j := j\sqrt{\log n}$.

For any j and every integer $i \in S_j := [(j-1)^2, j^2]\log n$, define a sequence of intervals,

$$I_j^i := \frac{i}{n} + \frac{L_j}{n}[-1,1].$$

Then for any $i \in S_j$ and $p_x \in I_j^i \cap I_j$, the corresponding Poisson probability satisfies

$$\begin{aligned} e^{-np_x}\frac{(np_x)^i}{i!} &= e^{-i}\frac{i^i}{i!}\cdot\exp\left(-(np_x-i)+i\cdot\log\left(1+\frac{np_x-i}{i}\right)\right) \\ &\geq \frac{1}{3\sqrt{i}}\cdot\exp\left(-\frac{2i}{3}\cdot\left(\frac{np_x-i}{i}\right)^2\right) \\ &\geq \frac{1}{9\sqrt{i}} \geq \frac{1}{9L_j}. \end{aligned}$$

On the other hand, the following upper bound holds.

$$\begin{aligned} e^{-np_x}\frac{(np_x)^i}{i!} &= e^{-i}\frac{i^i}{i!}\cdot\exp\left(-(np_x-i)+i\cdot\log\left(1+\frac{np_x-i}{i}\right)\right) \\ &\le e^{-i}\frac{i^i}{i!}\le\frac{1}{\sqrt{2\pi i}}\le\frac{1}{2L_j}. \end{aligned}$$

In other words, for any $p_x, i/n\in I_j$ that differ by at most L_j/n,

$$\Pr(\mathrm{Poi}(np_x)=i)\in\frac{1}{L_j}\left[\frac{1}{9},\frac{1}{2}\right].$$

Partition I_j into sub-intervals of equal length L_j/n. The partition has a size of at most $2\sqrt{\log n}$. Assign each probability $p_x\in I_j$ a length-L_j/n interval I_{p_x} centered at p_x. Then, each interval I_{p_x} covers at least one of the sub-intervals in the partition. Since there are exactly p_{I_j} intervals I_{p_x}, one can find a partition sub-interval I_j^s contained in at least $p_{I_j}/(2\sqrt{\log n})$ of them. Denote by $\mathcal{X}_s$ the collection of symbols corresponding to these intervals.

Next, we bound from below the entropy of the truncated profile $(\varphi_i)_{i\in nI_j^s}$ over nI_j^s. Denote by j_s the left end point of nI_j^s. By the chain rule of entropy for multiple random variables,

$$H((\varphi_i)_{i\in nI_j^s})=\sum_{i=j_s}^{j_s+L_j-1}H(\varphi_i|\varphi_{j_s},\ldots,\varphi_{i-1}).$$

Consider a particular term on the right-hand side with $i\in[j_s,j_s+L_j-1]$. By the conditional independence and fact that conditioning reduces entropy,

$$\begin{aligned} H(\varphi_i|\varphi_{j_s},\ldots,\varphi_{i-1}) &\ge H(\varphi_i|\varphi_{j_s},\ldots,\varphi_{i-1};\mathbb{1}_{j_s\le\mu_x\le i-1},x\in\mathcal{X}) \\ &= H(\varphi_i|\mathbb{1}_{j_s\le\mu_x\le i-1},x\in\mathcal{X}) \\ &= H(\varphi_i|\mathbb{1}_{j_s\le\mu_x\le i-1},x\in\mathcal{X}_s;\mathbb{1}_{j_s\le\mu_x\le i-1},x\notin\mathcal{X}_s) \end{aligned}$$

To characterize the condition, we define a random variable

$$K_i^s := \sum_{x \in \mathcal{X}_s} \mathbb{1}_{j_s \le \mu_x \le i-1}.$$

Note that $\mathbb{E}[\mathbb{1}_{j_s \le \mu_x \le i-1}] = \sum_{t=j_s}^{i-1} \Pr(\mathrm{Poi}(np_x) = t) \le (i-j_s)/(2L_j)$, which is at most $1/10$ for $i \le j_s + L_j/5$. The following lemma transforms this into a high-probability statement.

Lemma 3. *Let $Y_i, i \in [1,m]$ be independent indicator random variables. Let $Y := \sum_i Y_i$ denote their sum and $\lambda := \mathbb{E}[Y]$ denote the expected sum. Then for $c > 0$, we have*

$$\Pr(Y \ge \lambda(1+c)) \le \exp(-\lambda c^2/(2+2c/3)).$$

Below we consider only $i \le j_s + L_j/5$. Note that $c/(2+2c/3)$ is increasing for $c > 0$. Since $\mathbb{E}[K_i^s] = \sum_{x \in \mathcal{X}_s} \mathbb{E}[\mathbb{1}_{j_s \le \mu_x \le i-1}] \le |\mathcal{X}_s|/10$,

$$\Pr(K_i^s \ge |\mathcal{X}_s|/2) \le \exp(-36/35) < 1/2.$$

where we set $c = 4$ in the above lemma and assume that $|\mathcal{X}_s| \ge 3$ (assuming only $|\mathcal{X}_s| \ge 1$, the upper bound becomes $3/4$). Recall that

$$\begin{aligned} H(\varphi_i | \varphi_{j_s}, \ldots, \varphi_{i-1}) &\ge H(\varphi_i | \mathbb{1}_{j_s \le \mu_x \le i-1}, x \in \mathcal{X}_s; \mathbb{1}_{j_s \le \mu_x \le i-1}, x \notin \mathcal{X}_s) \\ &= \sum_{(c_x)_{x \in \mathcal{X}} \in \{0,1\}^{\mathcal{X}}} H(\varphi_i | \mathbb{1}_{j_s \le \mu_x \le i-1} = c_x, x \in \mathcal{X}_s) \\ &\qquad \times \Pr(\mathbb{1}_{j_s \le \mu_x \le i-1} = c_x, x \in \mathcal{X}_s). \end{aligned}$$

Denote by $V_s \subseteq \{0,1\}^{\mathcal{X}}$ the collection of $(c_x)_{x \in \mathcal{X}}$ satisfying $\sum_{x \in \mathcal{X}_s} c_x < |\mathcal{X}_s|/2$. The above derivation shows that

$$\sum_{(c_x)_{x \in \mathcal{X}} \in V_s} \Pr(\mathbb{1}_{j_s \le \mu_x \le i-1} = c_x, x \in \mathcal{X}_s) \ge \frac{1}{2}.$$

By independence, for any $(c_x)_{x\in\mathcal{X}} \in V_s$, we have

$$\begin{aligned}(\varphi_i|\mathbb{1}_{j_s\le\mu_x\le i-1} = c_x, x\in\mathcal{X}_s) &= \sum_{x\in\mathcal{X}:c_x=0} (\mathbb{1}_{\mu_x=i}|\mathbb{1}_{j_s\le\mu_x\le i-1}=0)\\ &= \sum_{x\in\mathcal{X}_s:c_x=0} (\mathbb{1}_{\mu_x=i}|\mathbb{1}_{j_s\le\mu_x\le i-1}=0)\\ &\qquad + \sum_{x\notin\mathcal{X}_s:c_x=0} (\mathbb{1}_{\mu_x=i}|\mathbb{1}_{j_s\le\mu_x\le i-1}=0).\end{aligned}$$

For any $x\in\mathcal{X}_s$ with $c_x = 0$, the corresponding indicator variable satisfies

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\mu_x=i}|\mathbb{1}_{j_s\le\mu_x\le i-1}=0] &= \frac{\Pr(\mathbb{1}_{\mu_x=i} \text{ and } \mu_x\notin[j_s,i-1])}{\Pr(\mu_x\notin[j_s,i-1])}\\ &= \frac{\Pr(\mathbb{1}_{\mu_x=i})}{1-\Pr(\mu_x\in[j_s,i-1])}\\ &= \frac{\frac{1}{L_j}\left[\frac{1}{9},\frac{1}{2}\right]}{1-\left[0,\frac{L_j}{5}\right]\cdot\frac{1}{L_j}\left[\frac{1}{9},\frac{1}{2}\right]}\\ &= \frac{1}{L_j}\left[\frac{1}{9},\frac{5}{9}\right].\end{aligned}$$

On the other hand, for any $x\notin\mathcal{X}_s$,

$$e^{-np_x}\frac{(np_x)^i}{i!} \le e^{-i}\frac{i^i}{i!} \le \frac{1}{\sqrt{2\pi i}} \le \frac{1}{2L_j}.$$

Therefore, the corresponding indicator variable satisfies

$$\mathbb{E}[\mathbb{1}_{\mu_x=i}|\mathbb{1}_{j_s\le\mu_x\le i-1}=0] = \frac{\Pr(\mathbb{1}_{\mu_x=i})}{1-\Pr(\mu_x\in[j_s,i-1])} \le \frac{\frac{1}{L_j}\left[0,\frac{1}{2}\right]}{1-\left[0,\frac{L_j}{5}\right]\cdot\frac{1}{L_j}\left[0,\frac{1}{2}\right]} \le \frac{5}{9}\cdot\frac{1}{L_j}.$$

To summarize, we have shown that $(\varphi_i|\mathbb{1}_{j_s\le\mu_x\le i-1}=c_x, x\in\mathcal{X}_s)$ is the sum of $|\mathcal{X}|$ independent Bernoulli random variables. Among these Bernoulli variables, at least $|\mathcal{X}_s|/2 \ge p_{I_j}/(2\sqrt{\log n})$ have a bias of $\frac{1}{L_j}\left[\frac{1}{9},\frac{5}{9}\right]$, while others have a bias of at most $\frac{5}{9}\cdot\frac{1}{L_j}$.

The following lemma, recently established by [HJ19], shows the relation among the entropy values of sums of independent Bernoulli random variables with different bias parameters.

Lemma 4. *Let $X_t, Y_t, t \in [m]$ be independent indicator random variables. Denote by X and Y the sums of X_t's and Y_t's, respectively. If $\mathbb{E}[X_t] \leq \mathbb{E}[Y_t] \leq 1/2, \forall t \in [m]$,*

$$H(\sum_t X_t) \leq H(\sum_t Y_t).$$

This lemma, together with the previous results, shows that

$$H(\varphi_i | \mathbb{1}_{j_s \leq \mu_x \leq i-1} = c_x, x \in \mathcal{X}_s) \geq H(\mathrm{bin}(p_{I_j}/(2\sqrt{\log n}), 1/(9L_j))).$$

The next lemma further bounds the entropy of a binomial random variable.

Lemma 5. *For any $m > 1$ and $q \in [1/m, 1-1/m]$,*

$$H(\mathrm{bin}(m,q)) \geq \frac{1}{2}\log\left((2\pi)^{1-(1-q)^m-q^m} mq(1-q)\right) - \frac{1}{12m}.$$

Proof. By definition, the left-hand side satisfies

$$\begin{aligned}
H(\mathrm{bin}(m,q)) &= -\sum_{t=0}^{m} \binom{m}{t} q^t (1-q)^{m-t} \log\left(\binom{m}{t} q^t (1-q)^{m-t}\right) \\
&= -\sum_{t=0}^{m} \binom{m}{t} q^t (1-q)^{m-t} (t\log q + (m-t)\log(1-q) \\
&\qquad\qquad + \log m! - \log t! - \log(m-t)!) \\
&= mH(\mathrm{Bern}(q)) - \log m! + \sum_{t=0}^{m} \binom{m}{t} q^t (1-q)^{m-t} (\log t! + \log(m-t)!).
\end{aligned}$$

By Stirling's formula, for any $t \geq 1$,

$$\log t! \geq \left(t + \frac{1}{2}\right)\log t + \frac{1}{2}\log(2\pi) - t.$$

Substituting the right-hand side into the above equation yields

$$S_m(q) := \sum_{t=0}^{m} \binom{m}{t} q^t (1-q)^{m-t} \log t! \geq \frac{1}{2}(1-(1-q)^m)\log(2\pi) - mq \\ + \sum_{t=1}^{m} \binom{m}{t} q^t (1-q)^{m-t} \left(t+\frac{1}{2}\right) \log t.$$

Let $g(x) := 0$ for $x \in [0,1)$ and $g(x) := (x+1/2)\log x$ for $x \geq 1$. Simple calculus shows that the function is concave. Applying the concavity of g to the last sum yields

$$\sum_{t=1}^{m} \binom{m}{t} q^t (1-q)^{m-t} \left(t+\frac{1}{2}\right) \log t \geq g\left(\sum_{t=0}^{m} \binom{m}{t} q^t (1-q)^{m-t} \cdot t\right) = \left(mq+\frac{1}{2}\right)\log(mq),$$

where the last step follows by the fact that $mq \geq 1$. A similar inequality holds for the weighted sum of $\log(m-t)!$. Consolidating these inequalities, we obtain

$$\begin{aligned} S_m(q) + S_m(1-q) &\geq \left(mq+\frac{1}{2}\right)\log(mq) + \left(m(1-q)+\frac{1}{2}\right)\log(m(1-q)) \\ &\quad + \frac{1}{2}(1-(1-q)^m)\log(2\pi) - mq + \frac{1}{2}(1-q^m)\log(2\pi) - m(1-q) \\ &= (m+1)\log m - mH(\mathrm{Bern}(q)) + \frac{1}{2}\log(q(1-q)) \\ &\quad + \frac{1}{2}(2-(1-q)^m - q^m)\log(2\pi) - m. \end{aligned}$$

On the other hand, for the $\log m!$ term,

$$\log m! \leq \left(m+\frac{1}{2}\right)\log m + \frac{1}{2}\log(2\pi) - m + \frac{1}{12m}.$$

Substituting the previous term bounds into the $H(\mathrm{bin}(m,q))$ expression yields

$$\begin{aligned} H(\mathrm{bin}(m,q)) &= mH(\mathrm{Bern}(q)) - \log m! + S_m(q) + S_m(1-q) \\ &\geq \frac{1}{2}\log\left((2\pi)^{1-(1-q)^m-q^m} mq(1-q)\right) - \frac{1}{12m}. \end{aligned}$$

□

Before continuing, we remark that the bound in the above lemma has the right dependence on $mq(1-q)$ in the sense that if we fix q and increase m, the lower bound converges to $\frac{1}{2}\log(\Theta(mq(1-q)))$. Another point to mention is that the above bound covers $q \in [1/m, 1-1/m]$, while Lemma 6 appearing later in this section covers $q \notin [1/m, 1-1/m]$. Note that the dependence on $mq(1-q)$ changes from logarithmic to linear, showing an "elbow effect" around $1/m$.

Assume that $p_{I_j}/(2\sqrt{\log n}) \geq 9L_j$, then for any $(c_x)_{x\in\mathcal{X}} \in V_s$,

$$H(\varphi_i | \mathbb{1}_{j_s \leq \mu_x \leq i-1} = c_x, x \in \mathcal{X}_s) \geq H(\mathrm{bin}(p_{I_j}/(2\sqrt{\log n}), 1/(9L_j)) \geq \frac{1}{2}.$$

Consolidating this with the previous results yields that

$$H(\varphi_i | \varphi_{j_s}, \ldots, \varphi_{i-1}) \geq \sum_{(c_x)_{x\in\mathcal{X}} \in V_s} \frac{1}{2} \cdot \Pr(\mathbb{1}_{j_s \leq \mu_x \leq i-1} = c_x, x \in \mathcal{X}_s) \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

where we utilize $p_{I_j}/(2\sqrt{\log n}) \geq 9L_j \geq 9$ and $(1-q)^m + q^m < 1/e$ for $\forall m \geq 3, q \in [1/m, 1/2]$. We can then bound the quantity of interest as follows.

$$\begin{aligned}
H((\varphi_i)_{i\in nI_j^s}) &= \sum_{i=j_s}^{j_s+L_j-1} H(\varphi_i | \varphi_{j_s}, \ldots, \varphi_{i-1}) \\
&\geq \sum_{i=j_s}^{j_s+L_j/5} H(\varphi_i | \varphi_{j_s}, \ldots, \varphi_{i-1}) \\
&\geq \frac{L_j}{5} \cdot \frac{1}{4} = \frac{L_j}{20} \\
&= \frac{1}{20\sqrt{\log n}} \min\left\{p_{I_j}, j \cdot \log n\right\}.
\end{aligned}$$

On the other hand, if $9L_j \geq p_{I_j}/(2\sqrt{\log n}) \gg 1$, we can further "compress" the truncated profile $(\varphi_i)_{i\in nI_j^s}$ over nI_j^s to reduce the effective value of L_j. Specifically, for any integer $t < L_j$, we define

the t-compressed version of $(\varphi_i)_{i\in nI_j^s}$ as

$$(\varphi_i)_{i\in nI_j^s}^t := \left(\sum_{i=j_s+(\ell-1)t}^{j_s+\ell t-1} \varphi_i\right)_{\ell\in[L_j/t]}.$$

Note that for each t, the length of $(\varphi_i)_{i\in nI_j^s}^t$ is $L_j^t := L_j/t$. For each entry in the compressed version, we can again express the entry as the sum of independent indicator random variables. Specifically,

$$\sum_{i=j_s+(\ell-1)t}^{j_s+\ell t-1} \varphi_i = \sum_{x\in\mathcal{X}} \mathbb{1}_{\mu_x\in[j_s+(\ell-1)t,\, j_s+\ell t-1]}.$$

Furthermore, for any $x\in\mathcal{X}_s$, the expectation of each indicator variable satisfies

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\mu_x\in[j_s+(\ell-1)t,\, j_s+\ell t-1]}] &= \sum_{i=j_s+(\ell-1)t}^{j_s+\ell t-1} e^{-np_x}\frac{(np_x)^i}{i!}\\ &= \frac{t}{L_j}\left[\frac{1}{9},\frac{1}{2}\right] = \frac{1}{L_j^t}\left[\frac{1}{9},\frac{1}{2}\right].\end{aligned}$$

Similarly, for any $x\in\mathcal{X}$, we have $\mathbb{E}[\mathbb{1}_{\mu_x\in[j_s+(\ell-1)t,\, j_s+\ell t-1]}] \le 1/(2L_j^t)$.

Now, choose t large enough so that $18L_j^t \ge p_{I_j}/(2\sqrt{\log n}) \ge 9L_j^t$. Following the reasoning in the previous case shows that

$$H((\varphi_i)_{i\in nI_j^s}) \ge H((\varphi_i)_{i\in nI_j^s}^t) \ge \Omega\left(\frac{1}{\sqrt{\log n}}\min\left\{p_{I_j}, j\cdot\log n\right\}\right).$$

It remains to consider the case of $O(\sqrt{\log n}) \ge p_{I_j} \ge 1$, for which we adopt our previous analysis.

Again, partition I_j into sub-intervals of equal length L_j/n. Then, assign each probability $p_x\in I_j$ a length-L_j/n interval I_{p_x} centered at p_x. By construction, each interval I_{p_x} covers at least one of the sub-intervals in the partition. Redefine any of these covered sub-intervals as I_j^s. Denote by $\mathcal{X}_s$ the collection of symbols corresponding to the covering intervals.

Note that $O(\sqrt{\log n}) \ge p_{I_j} \ge |\mathcal{X}_s| \ge 1$. For any $i\in[j_s, j_s+L_j/5]$, the previous analysis

shows that

$$H(\varphi_i|\varphi_{j_s},\ldots,\varphi_{i-1}) \geq H(\text{bin}(|\mathcal{X}_s|,1/(9L_j))\cdot(1-3/4)\,.$$

We bound the right-hand side with the following lemma.

Lemma 6. *For any $m \geq 1$, and $q \leq \min\{1/2, 1/m\}$ or $q \geq \max\{1/2, 1-1/m\}$,*

$$H(\text{bin}(m,q)) \geq \frac{m}{4}\min\{q,1-q\} \geq \frac{1}{4}mq(1-q).$$

Proof. By symmetry, we need to consider only the case of $q \in [0,1/m]$.

$$\begin{aligned}
H(\text{bin}(m,q)) &\geq H(\mathbb{1}_{\text{bin}(m,q)\geq 1}) \\
&= H(((1-q)^m, 1-(1-q)^m)) \\
&\geq -(1-q)^m(m\log(1-q)) \\
&\geq -\frac{m}{4}\log(1-q) \\
&\geq \frac{m}{4}\cdot q.
\end{aligned}$$

□

Consolidating the lemma and the chain rule of entropy yields,

$$\begin{aligned}
H((\varphi_i)_{i\in nI_j^s}) &= \sum_{i=j_s}^{j_s+L_j-1} H(\varphi_i|\varphi_{j_s},\ldots,\varphi_{i-1}) \\
&\geq \sum_{i=j_s}^{j_s+L_j/5} H(\varphi_i|\varphi_{j_s},\ldots,\varphi_{i-1}) \\
&\geq \frac{L_j}{5}\cdot\frac{|\mathcal{X}_s|}{4\cdot 9\cdot L_j}\cdot\left(1-\frac{3}{4}\right) = \frac{|\mathcal{X}_s|}{720} \\
&= \Omega\left(\frac{1}{\sqrt{\log n}}\min\left\{p_{I_j}, j\cdot\log n\right\}\right).
\end{aligned}$$

Alternatively, we can use the fact that adding independent random variables does not

decrease entropy, i.e., $H(Y+Z) \geq H(Y)$ for any independent variables Y and Z. Note that

$$(\varphi_i)^t_{i\in nI^s_j} = \sum_{x\in\mathcal{X}} (\mathbb{1}_{\mu_x=i})_{i\in I^s_j}.$$

Let y be an arbitrary symbol that belongs to $\mathcal{X}_s$. Then,

$$H((\varphi_i)_{i\in nI^s_j}) \geq H((\varphi_i)^t_{i\in nI^s_j}) \geq H((\mathbb{1}_{\mu_y=i})_{i\in I^s_j}) \geq H((\mathbb{1}_{\mu_y=j_s}, \mathbb{1}_{\mu_y=j_s+1})).$$

By the previous derivations, both $\Pr(\mu_y = j_s)$ and $\Pr(\mu_y = j_s+1)$ belong to $\frac{1}{L_j}[1/9, 1/2]$. Hence,

$$H((\varphi_i)_{i\in nI^s_j}) \geq H\left(\text{Bern}\left(\frac{2}{11}\right)\right) \geq \frac{2}{5} = \Omega\left(\frac{1}{\sqrt{\log n}} \min\left\{p_{I_j}, j\cdot\log n\right\}\right).$$

Note that this argument does not apply to other cases, since

$$H((\mathbb{1}_{\mu_y=i})_{i\in I^s_j}) = O(\log L_j) = O(\log n),$$

while $\min\left\{p_{I_j}, j\cdot\log n\right\}$ can be as large as $\tilde{\Theta}(n^{1/3})$ in general.

The proof is complete upon noting that indices with $j = O(1)$ corresponds to a total contribution of at most $O(1)$ to $H^{\mathcal{S}}_n(p)$ and $H^{\mathcal{S}}_n(p) = \tilde{\Theta}(\mathbb{E}[\mathcal{D}(\varphi)]) = \tilde{\Theta}(D(\varphi))$, with probability at least $1 - O(1/\sqrt{n})$.

Summary The simple expression shows that $\mathcal{H}^s_n(p)$ characterizes the variability of ranges that the actual probabilities spread over. As Theorem 1 shows, $\mathcal{H}^s_n(p)$ closely approximates $E_n(p)$, the value around which $\mathcal{D}_n \sim p$ concentrates (Theorem 9) and $\mathcal{H}_n(p)$ lies (Thoerem 1). Henceforth, we use $\mathcal{H}^s_n(p)$ as a proxy for both $\mathcal{H}_n(p)$ and $\mathcal{D}_n$, and study its attributes and values.

Let $p \in \Delta$ be an arbitrary discrete distribution. Recall that in Section 5.A, we partition the

unit interval into a sequence of ranges,

$$I_j := \left((j-1)^2 \frac{\log n}{n},\ j^2 \frac{\log n}{n} \right], 1 \le j \le \sqrt{\frac{n}{\log n}},$$

denote by p_{I_j} the number of probabilities p_x belonging to I_j, and relate $E_n(p)$ to an induced shape-reflecting quantity,

$$\mathcal{H}_n^s(p) := \sum_{j \ge 1} \min \left\{ p_{I_j}, j \cdot \log n \right\},$$

the sum of the effective number of probabilities lying within each range.

The simple expression of $\mathcal{H}_n^s(p)$ shows that it characterizes the variability of ranges the actual probabilities spread over. As Theorem 1 shows, $\mathcal{H}_n^s(p)$ closely approximates $E_n(p)$, the value around which $\mathcal{D}_n \sim p$ concentrates (Theorem 9) and $\mathcal{H}_n(p)$ lies (Thoerem 1). In this section, we use $\mathcal{H}_n^s(p)$ as a proxy for both $\mathcal{H}_n(p)$ and $\mathcal{D}_n$, and study its attributes and values.

To further our understanding of profile entropy and dimension, in the next two sections, we investigate the analytical attributes of $\mathcal{H}_n^s(p)$ concerning monotonicity and Lipschitzness.

5.A.4 Extension: Profile Entropy Estimation via Monotonicity

Among the many attributes that $\mathcal{H}_n^s(p)$ possesses, monotonicity is perhaps most intuitive. One may expect a larger value of $\mathcal{H}_n^s(p)$ as the sample size n increases, since additional observations reveal more information about the variability of probabilities. Below we confirm this intuition.

Theorem 12. *For any $n \ge m \gg 1$ and $p \in \Delta$,*

$$\mathcal{H}_n^s(p) \ge H_m^{\mathcal{S}}(p).$$

The above result that lowerly bounds $\mathcal{H}_n^s(p)$ with $H_m^{\mathcal{S}}(p)$ for $m \le n$. Besides this, a more desirable result is to upperly bound $\mathcal{H}_n^s(p)$ with some function of $H_m^{\mathcal{S}}(p)$. Such a result

will enable us to draw a sample of size $m \le n$, obtain an estimate of $H_m^{\mathcal{S}}(p)$ from $\mathcal{D}_m$ (by the entropy-dimension equivalence), and use it to bound the value of $\mathcal{H}_n^s(p)$ for a much larger sample size n.

With such an estimate, we can perform numerous tasks such as *predicting* the performance of PML when more observations are available, or the space needed for storing the profile of a longer sample sequence. These applications are closely related to the recent works on *learnability estimation* by [KV18, KVB19], namely, one wish to know how many (additional) observations are required for a learning algorithm to achieve a certain level of performance.

The next theorem provides a simple and tight upper bound on $\mathcal{H}_n^s(p)$ in terms of $H_m^{\mathcal{S}}(p)$.

Theorem 13. *For any $n \ge m \gg 1$ and $p \in \Delta$,*

$$\mathcal{H}_n^s(p) \le \sqrt{\frac{n \log n}{m \log m}} \cdot H_m^{\mathcal{S}}(p).$$

Estimation Before continuing to the proof, we present some direct implications.

1. If for $m = \Omega(n^{0.01})$, we have $H_m^{\mathcal{S}}(p) \ll \sqrt{m}$, then $H_n^{\mathcal{S}}(p) \ll \sqrt{n}$.

2. For any two integers $m \le n$ and distribution p,

$$\frac{H_m^{\mathcal{S}}(p)}{\sqrt{m \log m}} \ge \frac{H_n^{\mathcal{S}}(p)}{\sqrt{n \log n}}.$$

   In other words, the sequence $A_m := H_m^{\mathcal{S}}(p)/\sqrt{m \log m}$, $m \le n$, is monotonically decreasing and converges to A_n. As we increase the value of m, $(\sqrt{n \log n} \cdot A_m)$, which can be viewed as our estimate of $H_n^{\mathcal{S}}(p)$, is getting more and more accurate. For the purpose of adaptive estimation, if $n = 2^t$, we can choose $m = 2^0, 2^1, \ldots, 2^t$.

Proof. Below we prove both the lower and upper bounds. For clarity, denote by $p(m, j)$ the value of p_{I_j} corresponding to $H_m^{\mathcal{S}}(p)$, and $p(n, j)$ the value of p_{I_j} corresponding to $H_n^{\mathcal{S}}(p)$. Furthermore,

denote $r := \sqrt{(n/m)((\log m)/\log n)}$, which is treated as an integer. Then, by the definition of $H^{\mathcal{S}}_{\cdot}$,

$$\begin{aligned}
rH^{\mathcal{S}}_m(p) &= r\sum_{j\geq 1}\min\{p(m,j), j\cdot\log m\} \\
&= \sum_{j\geq 1}\min\left\{r\cdot\sum_{i=rj-r+1}^{rj}p(n,i),\ rj\cdot\log m\right\} \\
&\geq \sum_{j\geq 1}\sum_{t=0}^{r-1}\min\left\{\sum_{i=rj-r+1}^{rj}p(n,i),\ (rj-t)\cdot\log m\right\} \\
&\geq \sum_{j\geq 1}\sum_{t=0}^{r-1}\min\{p(n,rj-t),\ (rj-t)\cdot\log m\} \\
&= \sum_{i\geq 1}\min\{p(n,i),\ i\cdot\log m\} \\
&\geq \frac{\log m}{\log n}\cdot H^{\mathcal{S}}_n(p).
\end{aligned}$$

The lower-bound part basically follows by reversing the above inequalities.

$$\begin{aligned}
H^{\mathcal{S}}_n(p) &= \sum_{i\geq 1}\min\{p(n,i),\ i\cdot\log n\} \\
&= \sum_{j\geq 1}\sum_{t=0}^{r-1}\min\{p(n,rj-t),\ (rj-t)\cdot\log n\} \\
&\geq \sum_{j\geq 1}\sum_{t=0}^{r-1}\min\{p(n,rj-t),\ (rj-r+1)\cdot\log n\} \\
&\geq \sum_{j\geq 1}\min\left\{\sum_{t=0}^{r-1}p(n,rj-t),\ (rj-r+1)\cdot\log n\right\} \\
&= \sum_{j\geq 1}\min\{p(m,j),(rj-r+1)\cdot\log m\} \\
&\geq H^{\mathcal{S}}_m(p).
\end{aligned}$$

This completes the proof of the theorem. □

5.A.5 Extension: Lipschitzness of Profile Entropy

Note that we can view $\mathcal{H}_n^S(p)$ as a distribution property. In this section, we establish the Lipschitzness of $\mathcal{H}_n^S(p)$ under a weighted Hamming distance and the ℓ_1 distance between distributions. Precisely, given two distributions $p, q \in \Delta$, the vanilla *Hamming distance* is

$$h(p,q) := \sum_x \mathbb{1}_{p_x \neq q_x}.$$

This may not be suitable for the purpose of statistical inference since the two distributions could differ at many symbols, while these symbols account for only a negligible total probability and has little effects on most induced statistics. To address this, we propose a *weighted Hamming distance*

$$h_w(p,q) := \sum_{x \in \mathcal{X}} \max\{p_x, q_x\} \cdot \mathbb{1}_{p_x \neq q_x}.$$

The next result measures the Lipschitzness of H_n^S under h_w.

Theorem 14. *For any integer n, and distributions p and q, if $h_w(p,q) \leq \varepsilon$ for some $\varepsilon \geq 1/n$,*

$$\left| \mathcal{H}_n^S(p) - H_n^S(q) \right| \leq \tilde{O}(\sqrt{\varepsilon n}).$$

Proof. Recall that the quantity of interest is

$$\mathcal{H}_n^S(p) := \sum_{j \geq 1} \min\left\{ p_{I_j}, j \cdot \log n \right\}.$$

Given the bound of $h_w(p,q) \leq \varepsilon$, we denote by $\mathcal{Y}$ the collection of symbols x at which $p_x \neq q_x$. By definition, we have both $\sum_{x \in \mathcal{Y}} p_x \leq \varepsilon$ and $\sum_{x \in \mathcal{Y}} q_x \leq \varepsilon$. Below, we show that these symbols modify the value of $\mathcal{H}_n^S(p)$ by at most $\tilde{O}(\sqrt{\varepsilon n})$. By symmetry, the same claim also holds for the distribution q. Combining the two claims yields the desired result.

First, we consider $x \in \mathcal{Y}$ satisfying $p_x = 0$ or $p_x \in I_1 = (0, (\log n)/n]$. Such a symbol either

does not contribute the value of $\mathcal{H}_n^s(p)$, or affects only the value of the first term $\min\{p_{I_1}, \log n\}$, which is at most $\log n$. Hence the claim holds for this case.

Next, consider symbols $x \in \mathcal{Y}$ satisfying $p_x \in I_j = ((j-1)^2 \frac{\log n}{n},\ j^2 \frac{\log n}{n}]$ for some $j \geq 2$ and denote the collection of them by $\mathcal{Z} \subseteq \mathcal{Y}$. By the above assumption, we have $\sum_{x \in \mathcal{Z}} p_x \leq \varepsilon$. To maximize their impact on $\mathcal{H}_n^s(p)$ under this constraint, we should set their values to be

$$p_j := (j-1)^2 \frac{\log n}{n},\ j = 2, \ldots J,$$

for some J to be determined, where each p_j repeats exactly $j \log n$ times. Then, the symbols in $\mathcal{Z}$ contributes at most $\sum_{j=2}^{J} j \log n = (\log n)(J-1)(J+2)/2$ to $\mathcal{H}_n^s(p)$, and the above constraint on the total probability mass bounds transforms to

$$\varepsilon \geq \sum_{x \in \mathcal{Z}} p_x \geq \sum_{j=2}^{J} (j \log n) \cdot (j-1)^2 \frac{\log n}{n} \geq \frac{(\log n)^2}{12n} J(J^2-1)(-2+3J).$$

Therefore in this case, the contribution is again $\tilde{O}(\sqrt{\varepsilon n})$, which completes the proof. □

Replacing $\max\{p_x, q_x\}$ with $|p_x - q_x|$ induces a common similarity measure, the ℓ_1 distance. The next theorem is an analog to Theorem 14 under this classical distance.

Theorem 15. *For any integer n, and distributions p and q, if $\ell_1(p,q) \leq \varepsilon$ for some $\varepsilon \geq 0$,*

$$\left| \mathcal{H}_n^s(p) - c H_n^{\mathcal{S}}(q) \right| \leq O((\varepsilon n)^{2/3}),$$

where c is a constant in $[1/3, 3]$. Note that the inequality is significant iff $\varepsilon \leq \tilde{\Theta}(1/n^{1/4})$, since the value of $\mathcal{H}_n^s(p)$ is at most $O(\sqrt{n \log n})$ for all p.

By symmetry, it suffices to prove that under the conditions in Theorem 15,

$$H_n^{\mathcal{S}}(p) \leq 3 H_n^{\mathcal{S}}(q) + O((\varepsilon n)^{2/3}).$$

Proof. Consider the optimization problem of modifying p by at most ε and maximizing the increase in $H_n^{\mathcal{S}}(p)$. For each j and each probability $p_x \in j$, denote by p'_x the modified value. Depending on the location of p'_x, there are three types of possible modifications, as illustrated below.

- For the first type, we still have $p'_x \in I_j$. This does not change the value of p_{I_j} and hence does not increase $H_n^{\mathcal{S}}(p)$.

- For the second type, we have $p'_x \in I_{j-1}$ or $p'_x \in I_{j+1}$. If $p_{I_j} \leq j \cdot \log n$, this will decrease the value of $\min\{p_{I_j}, j \cdot \log n\}$ by 1 and increase the value of $\min\{p_{I_{j-1}}, (j-1) \cdot \log n\}$ or $\min\{p_{I_{j+1}}, (j+1) \cdot \log n\}$ by at most one. Hence in this case, the value of $H_n^{\mathcal{S}}(p)$ can only decrease. If $p_{I_j} > j \cdot \log n$, then $\min\{p_{I_j}, j \cdot \log n\} = j \cdot \log n$. For a particular j, all such modifications can increase the value of $H_n^{\mathcal{S}}(p)$ by at most $(j-1)\log n + (j+1)\log n = 2j \log n$, which is twice the value of $\min\{p_{I_j}, j \cdot \log n\}$. Hence, all such modifications, when combined, increase the value of $H_n^{\mathcal{S}}(p)$ by at most $2H_n^{\mathcal{S}}(p)$.

- For the third type, we have $p'_x \in I_i$ and $|i-j| \geq 2$. If $i < j$, we require a probability mass of at least $((j-1)^2 \log n - i^2 \log n)/n \geq (i \log n)/n$, where $j \geq 3$. If $i > j$, we require a probability mass of at least $((i-1)^2 \log n - j^2 \log n)/n \geq (i \log n)/n$. The number of such modifications that could lead to an increase in the value of $H_n^{\mathcal{S}}(p)$ is at most $i \log n$. For each i, let c_i denote the number of such modifications that will lead to an increase of $H_n^{\mathcal{S}}(p)$. Then, the total increase is $\sum_i c_i$, each c_i is at most $i \log n$, and the total required probability mass required is at least $\sum_i c_i \cdot (i \log n)/n \leq \varepsilon$.

  Let $\{c_i\}$ be the optimal solution that maximizes $\sum_i c_i$. Assume that there are two indices $i < j$ satisfying $c_i < i \log n$ and $c_j > 0$. Then, if we replace c_i and c_j by $c_i + 1$ and $c_j - 1$, respectively, $\sum_i c_i$ will not change and $\sum_i c_i \cdot (i \log n)/n$ will decrease. Hence, we can assume that there exists i' satisfying $c_i = i \log n, \forall i < i'$ and $c_i = 0, \forall i > i'$. In addition, assuming

$\varepsilon n \geq \log n$ implies that $i' \geq 2$. Hence, we have $\sum_i c_i \leq (\log n) i'(i'+1)/2$ and

$$\sum_i c_i \leq 3.5 \cdot \left(\frac{n\varepsilon}{\sqrt{\log n}} \right)^{2/3}. \qquad \square$$

5.B Competitive-Optimal Property Inference

5.B.1 Theorem 3: Sufficiency of Profiles

Numerous practical applications call for inferring *property values* of an unknown distribution from its samples, such as entropy for graphical modeling [KF09], Rényi entropy for sequential decoding [Ari96], and support size for species richness estimation [Mag13]. Therefore, *property inference* has attracted considerable attention over the past few decades.

Property inference Formally, a *distribution property* over some collection $\mathrm{P} \subseteq \Delta$ is a functional $f : \mathrm{P} \to \mathbb{R}$ that associates with each distribution a real value. Given a sample X^n from an unknown distribution $p \in \mathrm{P}$, the problem of interest is to infer the value of $f(p)$. For this purpose, we employ another functional $\hat{f} : \mathcal{X}^* \to \mathbb{R}$, an *estimator* mapping every sample to a real value. We measure the statistical efficiency of $\hat{f}$ in approximating f over P by its *absolute error* $|\hat{f}(X^n) - f(p)|$.

Given $X^n \sim p \in \mathrm{P}$, the *minimal absolute error rate*, or simply *error*, that $\hat{f}$ achieves with probability at least $9/10$ is $r_n(p, \hat{f}) := \min\{r : \Pr(|\hat{f}(X^n) - f(p)| \leq r) \geq 9/10\}$, where the dependence on f is *implicit*. While p is often unknown, the *worst-case error* of an estimator $\hat{f}$ over all distributions in P is $r_n(\mathrm{P}, \hat{f}) := \max_{p \in \mathrm{P}} r_n(p, \hat{f})$, and the lowest worst-case error for P, achieved by the optimal estimator, is the *minimax error* $r_n(\mathrm{P}) := \min_{\hat{f}'} r_n(\mathrm{P}, \hat{f}')$.

Symmetric properties An important class of properties is the collection of symmetric ones, which encompasses numerous well-known distribution characteristics, such as Shannon entropy, Rényi entropy, support size, and ℓ_1 distance to the uniform distribution. Symmetry connects the estimation of such property to the sample profile, a sufficient statistic for the task in

hand. The general principle of maximum likelihood then provides an intuitive estimator, *profile maximum likelihood (PML)* [OSVZ04], that maximizes the probability of observing the profile.

An estimator is *profile-based* if its values depends on only the profile. The theorem below shows that profile-based estimators are sufficient for inferring symmetric properties.

Theorem 3 (Sufficiency of profiles)**.** *For any symmetric property f and set $\mathrm{P} \subseteq \Delta$, and estimator $\hat{f}$, we can construct an explicit estimator $\hat{F}$ over length-n profiles satisfying*

$$r_n(p, \hat{f}) = r_n(\mathrm{P}, \hat{F} \circ \varphi),$$

where both estimators can have independent randomness.

Proof. First we show that given estimator $\hat{f}$, there is an estimator $\hat{f}_s$ which is symmetric, i.e., invariant with respect to domain-symbol permutations, and achieves the same guarantee. To see this, consider a random permutation $\tilde{\sigma}$ chosen uniformly randomly from the collection of permutations over the underlying alphabet. Let $\hat{f}_s := \hat{f} \circ \tilde{\sigma}$. Then for any $p \in \mathcal{P}$,

$$\begin{aligned}
\Pr_{X^n \sim p}\left(\left|\hat{f}_s(X^n) - f(p)\right| > \varepsilon\right) &\overset{(a)}{=} \Pr_{X^n \sim p}\left(\left|\hat{f} \circ \tilde{\sigma}(X^n) - f(p)\right| > \varepsilon\right) \\
&\overset{(b)}{=} \sum_{\sigma} \Pr_{X^n \sim p}\left(\left|\hat{f} \circ \sigma(X^n) - f(p)\right| > \varepsilon \,\middle|\, \tilde{\sigma} = \sigma\right) \cdot \Pr\left(\tilde{\sigma} = \sigma\right) \\
&\overset{(c)}{=} \sum_{\sigma} \Pr_{X^n \sim p}\left(\left|\hat{f} \circ \sigma(X^n) - f(p)\right| > \varepsilon\right) \cdot \Pr\left(\tilde{\sigma} = \sigma\right) \\
&\overset{(d)}{=} \sum_{\sigma} \Pr_{X^n \sim \sigma(p)}\left(\left|\hat{f}(X^n) - f(\sigma(p))\right| > \varepsilon\right) \cdot \Pr\left(\tilde{\sigma} = \sigma\right) \\
&\overset{(e)}{<} \sum_{\sigma} \delta \cdot \Pr\left(\tilde{\sigma} = \sigma\right) \\
&\overset{(f)}{=} \delta,
\end{aligned}$$

where (a) follows by the definition of $\hat{f}_s$; (b) follows by the law of total probability; (c) follows by the independence between $\tilde{\sigma}$ and X^n; (d) follows by the symmetry of f and the equivalence of

applying σ to X^n and to p; (e) follows by the fact that $\sigma(p) \in \mathcal{P}$ and the guarantee satisfied by the estimator $\hat{f}$; and (f) follows by the law of total probability.

Before we proceed further, we introduce the following definitions. For any sequence x^n, the *sketch* of a symbol x in x^n is the set of indices $i \in [n]$ for which $x_i = x$. The *type* of a sequence x^n is the set $\tau(x^n)$ of sketches of symbols appearing in x^n.

Since $\hat{f}_s$ is symmetric, there exists a mapping $\hat{f}_\tau$ over types satisfying $\hat{f}_s = \hat{f}_\tau \circ \tau$. Due to the i.i.d.assumption on the sample generation process, given the profile of a sample sequence, all the different types corresponding to this profile are equally likely. Let Λ be a mapping that recovers this relation, i.e., Λ maps each profile uniformly randomly to a type having this profile.

Then, for any $p \in \mathcal{P}$ and $X^n \sim p$,

$$\hat{f}_s(X^n) = \hat{f}_\tau \circ \tau(X^n) = \hat{f}_\tau \circ \Lambda \circ \varphi(X^n).$$

Consequently, the mapping $\hat{F} := \hat{f}_\tau \circ \Lambda$ is a profile-based estimator that satisfies

$$\Pr_{X^n \sim p}\left(\left|\hat{F}(\varphi(X^n)) - f(p)\right| > \varepsilon\right) = \Pr_{X^n \sim p}\left(\left|\hat{f}_s(X^n) - f(p)\right| > \varepsilon\right) < \delta,\ \forall p \in \mathcal{P}. \qquad \square$$

5.B.2 Theorem 4: Competitiveness of PML

Naturally and generally, we study symmetric property inference over a distribution collection $\mathrm{P} \subseteq \Delta$ that is also *symmetric*, i.e., if $p \in \mathrm{P}$, then P as well contains all the symbol-permuted versions of p. For every sample $x^n \in \mathcal{X}^n$ and symmetric P, the *PML estimator* over P maps x^n to a distribution

$$\mathcal{P}_\varphi(x^n) := \arg\max_{p \in \mathrm{P}} \Pr_{X^n \sim p}\left(\varphi(X^n) = \varphi(x^n)\right).$$

Given a sample $X^n \sim p \in \mathrm{P}$ and a symmetric property p, the PML plug-in estimator uses $f \circ \mathcal{P}(X^n)$ to estimate $f(p)$. Recent researches [ADOS17, HO19a] show that for an extensive family of symmetric properties, including the previously mentioned four, the PML plug-in

estimator *universally* achieves minimax error in the large-alphabet regime, up to constant factors.

The next result shows that the PML estimator is adaptive to the simplicity of underlying distributions in inferring all symmetric properties, over any symmetric P. Specifically, the theorem states that the n-sample PML plug-in essentially performs as well as the optimal $n/\mathcal{H}_n(p)$-sample estimator, which approaches the performance of the optimal n-sample estimator if p has a small $\mathcal{H}_n(p)$. Furthermore, for any property and estimator, there is a symmetric set P' for which this $1/\mathcal{H}_n(p)$ ratio is *optimal*.

Theorem 4 (Competitiveness of PML)**.** *For any symmetric property f and set $\mathrm{P} \subseteq \Delta$, and every distribution $p \in \mathrm{P}$, the PML plug-in estimator satisfies*

$$r_n\big(p, f \circ \mathcal{P}_\varphi\big) \le 2 r_{n_p}(\mathrm{P}),$$

where $n_p := n/\mathcal{H}_n(p)$. On the other hand, for any estimator $\hat{f}$ and symmetric property f, there exists a symmetric set $\mathrm{P}' \subseteq \Delta$ such that for some $p \in \mathrm{P}'$,

$$r_n(p, \hat{f}\,) \ge 2 r_{n_p}\big(\mathrm{P}'\big)\,.$$

5.B.3 Prior Work and Discussions

Results Recent years have shown interests in determining the limits of inferring symmetric distribution properties. Building upon worst-case analysis, the major contribution of these works is showing that for several specific properties, one can design more involved estimators whose worst-case performance is better than the empirical-distribution plug-in estimators (*empirical estimators*), over $\Delta_{\mathcal{X}}$ for some *finite* alphabet $\mathcal{X}$. Note that $\Delta_{\mathcal{X}}$ is a special symmetric distribution collection.

For example, the empirical estimator for Shannon entropy has a worst-case error rate of $\Theta(|\mathcal{X}|/n)$, whereas the minimax error rate is $\Theta(|\mathcal{X}|/(n\log n))$ [VV11a, VV13, JVHW15, WY16,

ADOS17, HO19a, HO19c, HO20a]. Similar results also hold for support size and ℓ_1 distance to the uniform distribution over $\mathcal{X}$ (See [VV11a, VV13, ADOS17, JHW16, WY19, HO19a, HO19c, HO20a]). One observation is that all these properties are in the form of $\sum_x f_r(p_x)$, where f_r is a relative smooth real function (for support size, one needs a lower bound like $1/|\mathcal{X}|$ on the positive probabilities, which effectively smoothes the function).

It is apparent that most symmetric properties are not in the $\sum_x f_r(p_x)$ form. A simple example is Rényi entropy, for which the learning error rates exhibit a significantly different behavior. Specifically, for a power parameter $\alpha > 1, \alpha \in \mathbb{N}$, the minimax error of inferring Rényi entropy varies according to $|\mathcal{X}|$ and n as follows [AOST16].

If $n \lesssim |\mathcal{X}|^{1-1/\alpha}$ (sample-sparse regime), then $r_n(\Delta_{\mathcal{X}}) \gtrsim \max_p f(p)$ (consistent estimation is impossible); if $n \gtrsim |\mathcal{X}|^{1+1/\alpha}$ (large-sample regime), then $r_n(\Delta_{\mathcal{X}}) \simeq (|\mathcal{X}|^{1-1/\alpha}/n)^{1/2}$, which is *achieved* by the empirical estimator (trivial regime); if $|\mathcal{X}|^{1-1/\alpha} \lesssim n \lesssim |\mathcal{X}|^{1+1/\alpha}$, then the empirical estimator has an order $\max\{|\mathcal{X}|/n, 1\}$ worst-case error, whereas the minimax error is $(|\mathcal{X}|^{1-1/\alpha}/n)^{1/2}$ (potentially much lower than that of empirical).

The recent work of [HO19a] significantly extends our understanding of symmetric property estimation by showing that the PML estimator is sample optimal for all $\sum_x f_r(p_x)$ properties that are approximately *Lipschitz*, and is as good as the best known estimators for Rényi entropy of power $\alpha > 3/4$. The paper also presents resulting on other tasks such as testing.

Given the special structures, even the combination of all the properties mentioned above corresponds to only an extremely small subclass of symmetric properties. The general landscape for how the worst-case error rate behaves when we consider either the empirical or the minimax estimator is far from understood, even for just $\Delta_{\mathcal{X}}$. In fact, even for Rényi entropy, a simple and widely studied property, the minimax rates are not fully characterized – the lower and upper bounds in [AOST16] for non-integer powers do not match in all parameter regimes. Ideally, there should be a set of formulas such that once the explicit form of f is available, the respective error rates can be computed, and more importantly, an explicit algorithm can be derived.

Our result pushes forward the general understanding of symmetric property estimation. It leverages the method of PML to derive competitive learning guarantees for all symmetric properties and distribution collections. The theorem even adapts itself to individual distributions, leading to numerous nontrivial estimation results without introducing sophisticated analysis or additional algorithms.

Methods As the task involves two components, the property and distribution (probability multiset), the design of statistical methods also advances in two veins.

The first vein concerns constructing a universal plug-in estimator for all *symmetric properties*. A symmetric property is invariant under symbol permutations, hence it suffices to obtain an accurate estimate of the probability multiset.

One method is PML, the approach that our theorem adopts. Recently, following the papers by [Das12, ADOS17], the work of [HO19a] shows that for any symmetric property that is in the form of $\sum_x f_r(p_x)$ and appropriately Lipschitz, both the profile maximum likelihood [OSVZ04] and its near-linear-time computable variant in [CSS19b] achieve the optimal sample complexity up to small constant factors.

Another method is moment matching via linear programming (LP). In typical works using LP, such as [VV11a, VV13, VV16, HJW18], one first estimates the (lower-order) moments of the underlying distributions (e.g., $\sum_x p_x^i$ for $i \leq \log n$), which are also symmetric properties, and then finds a distribution through an LP method (up to domain-symbol permutations), whose lower order moments match with the estimates. These methods are known to achieve the minimax error rates over $\Delta_{\mathcal{X}}$ for only a few specific properties, such as entropy, support size (also assume a $1/|\mathcal{X}|$ lower bound on the positive probabilities), and ℓ_1-distance to the uniform distribution.

The second vein of methods addresses the bias of empirical estimators and (often partially) replaces the given property by a bias-corrected polynomial, for which we can efficiently construct a near-unbiased estimator. There are mainly three different types of constructions for the bias-corrected polynomial: using classical minimax approximation [JVHW15, JHW16, WY16, WY19,

HO19c], applying smoothing techniques to the coefficients of the unbiased estimator [OSW16, HOSW18, HL20a], and computing the derivative of the (property's) Bernstein polynomial and employing the integral of its minimax approximation [HO20a].

Early works in this direction address specific properties, such as entropy [JVHW15, WY16], support size [WY19], support coverage [OSW16], and ℓ_1-distance to the uniform distribution [JHW16], and determine their respective minimax error rates. Recent works consider broader families of properties [HOSW18, HO19c, HO20a, HL20a], in particular those in the $\sum_x f_r(p_x)$ form and appropriately Lipschitz. Besides these results, some state-of-the-art Rényi entropy estimators [AOST16] also use polynomial approximation. Excluding properties in these special forms, it is unknown whether these techniques/methods work for the large amount of symmetric properties in general, even just over $\Delta_{\mathcal{X}}$.

Outline The rest of Appendix 5.B presents the proof of the our result on PML. For clarity, we divide the full proof into three parts: a) the sufficiency of profiles for estimating symmetric properties (already established above); b) the standard "median trick" often used to boost the confidence of learning algorithms; c) the PML method and its competitiveness to the min-max estimators. As one may expect, the proof utilizes several previously established results.

5.B.4 Proof of Theorem 4

Proof outline We begin with a proof sketch on the high level. While our theorem states only a constant-error-probability result for the vanilla PML, the guarantee holds for approximations of PML and any general error probability bound δ, and this outline corresponds to the general setting.

1 For simplicity, let k denote the (expected or high-probability) dimension of a length-n profile from an unknown $p \in \Delta$, and refer to the actual random quantity $\mathcal{D}_n \sim p$ as "dimension".

2 Let's say $p \in \mathcal{P}$ (which is symmetric), and we have an m-sample estimator over $\mathcal{P}$ with an

(ε, δ) guarantee, i.e., for every distribution in $\mathcal{P}$, the estimator learns its property value up to an ε error, with probability at least $1-\delta$. In addition, we assume that $m \ll n$ with the ratio $r := n/m$ to be determined.

3 Now, assume that r has been properly chosen, and we could utilize at most r copies of the m-sample estimator to construct an n-sample $(\varepsilon, \delta \cdot \exp(-2k))$ estimator (the existence of r follows by the standard "median trick"). Furthermore, by the sufficiency of profiles (Theorem 3), there is a profile-based estimator that achieves the same guaranty.

4 Divide all length-n profiles into two groups: one group with dimension at most of order k (hiding logarithmic factors), the other with dimension much larger than k.

5 By the concentration of sample profile dimensions (e.g., Theorem 9), the profile of an arbitrary sample from p belongs to the first group with high probability (say at least $1-1/n$), we can safely ignore the second group.

6.1 Pick a profile from "the first group", if its probability is $\gg \delta \cdot \exp(-k)$, the approximate PML (APML) will have a probability of $\gg \delta \cdot \exp(-2k)$. Here, the definition of APML is based on profile probabilities – for every length-n sample, its profile probability under the true distribution and the APML estimate should differ by a factor of at most $\exp(k)$ (more generally, a fixed factor of at least 1, which covers the vanilla PML). This definition is analogous to those in [ADOS17] and [CSS19a, CSS19b].

6.2 So, the *profile-based estimator* must work properly on both distributions, the original and the APML. Triangle inequality then relates the property values of these distributions (by eliminating the estimator's value) and yields a 2ε estimation guarantee for the APML.

7.1 On the other hand, if the profile we picked has a probability at most $\delta \cdot \exp(-k)$, then the APML may fail, i.e., not produce a reasonable estimate.

7.2 However, there are at most (ignore logarithmic factors in the exponent) $\exp(k)$ such profiles, hence by the union bound, the total probability of failing is at most $\delta + 1/n$.

8 Finally, we tune parameter r, which becomes something like k, up to logarithmic factors. Utilizing our entropy-dimension equivalence (Theorem 1) completes the proof.

Median Trick The following argument is standard method for boosting the confidence of learning algorithms, commonly known as the *median trick*.

Lemma 7 (Median trick)**.** *Let $\alpha, \beta \in (0,1)$ be real parameters satisfying $1/10 \geq \alpha > \beta$. For an accuracy $\varepsilon > 0$ and a distribution set $\mathcal{P} \subseteq \Delta$, if there exists an estimator $\hat{f}_A$ such that*

$$\Pr_{X^n \sim p}\left(\left|\hat{f}_A(X^n) - f(p)\right| > \varepsilon\right) < \alpha, \ \forall p \in \mathcal{P},$$

we can construct another estimator $\hat{f}_B$ that takes a sample of size $m := \left\lceil \frac{4n}{\log \frac{1}{2\alpha}} \log \frac{1}{\beta} \right\rceil$ and achieves

$$\Pr_{Y^m \sim p}\left(\left|\hat{f}_B(Y^m) - f(p)\right| > \varepsilon\right) < \beta, \ \forall p \in \mathcal{P}.$$

Proof. Given $t \in \mathbb{N}$ i.i.d.copies of $\hat{f}_A(X^n)$, the probability that less than half of them satisfy the inequality in the parentheses is at least

$$\Pr\left(\sum_{i=1}^{t} \mathbb{1}_{A_i} < \frac{t}{2} \text{ for } A_i\text{'s satisfying } \Pr(A_i) < \alpha\right) \geq \Pr\left(\mathrm{bin}\,(t,\alpha) < \frac{t}{2}\right).$$

By the law of total probability, the right-hand side equals to

$$\begin{aligned} 1 - \Pr\left(\mathrm{bin}\,(t,\alpha) \geq \frac{t}{2}\right) &\geq 1 - \exp\left(\left(\left(\frac{1}{2\alpha} - 1\right) - \frac{1}{2\alpha}\log\frac{1}{2\alpha}\right) \cdot \alpha t\right) \\ &\geq 1 - \exp\left(-\frac{t}{4}\log\frac{1}{2\alpha}\right), \end{aligned}$$

where the first step follows by the Chernoff bound of binomial random variables, and the second

step follows by $\alpha \le 1/10$ and the inequality $c - 1 - \frac{c}{2}\log c > 0, \forall c \ge 5$.

Set $t := \left\lceil \frac{4}{\log \frac{1}{2\alpha}} \log \frac{1}{\beta} \right\rceil$, the right-hand side is at least $1-\beta$.

Therefore, given a sample of size $m = t \cdot n$, we can partition it into t sub-samples of equal size, apply the estimator $\hat{f}_A$ to each subsample, and define the median of the corresponding estimates as $\hat{f}_B$.

By the previous reasoning, this estimator satisfies

$$\Pr_{Y^m \sim p}\left(\left|\hat{f}_B(Y^m) - f(p)\right| > \varepsilon\right) < \beta, \ \forall p \in \mathcal{P}. \qquad \square$$

Proof of the theorem. For any tolerance $\delta \in (0,1)$ and distribution $p \in \Delta$, define the (δ, n)-*typical cardinality of profiles with respect to* p as the smallest cardinality $C_{\delta,n}(p)$ of a set of length-n profiles such that the probability of observing a sample from p with a profile in this set is at least $1-\delta$. The following lemma provides a tight characterization of $C_{\delta,n}(p)$ in terms of the dimension of $\Phi^n \sim p$.

Lemma 8. *For any $p \in \Delta$ and $\Phi^n \sim p$, with probability at least $1 - 6/\sqrt{n}$,*

$$C_{\frac{6}{\sqrt{n}},n}(p) \le n^{8(\mathcal{D}_n + 20\log n)}.$$

The proof of the lemma follows by recursively applying Theorem 9. Specifically, let $d := 2E_n(p) + 3\log n$, which is at least $\mathcal{D}_n \sim p$, with probability at least $1 - 6/\sqrt{n}$. Then,

$$C_{\frac{6}{\sqrt{n}},n}(p) \le \binom{n}{d}\binom{n+d-1}{d-1} \le n^{2d-1} \le n^{2(2E_n(p)+3\log n)} \le n^{8\mathcal{D}(\Phi^n)+20\log n},$$

where the last inequality holds with with probability at least $1 - 6/\sqrt{n}$.

Now, let f be a symmetric property over $\mathcal{P}$. For simplicity, we will establish the theorem for the vanilla PML, since as our *proof outline* shows, the proof for any approximate PML (APML) is essentially the same. In addition, for a sequence x^n with profile $\phi := \varphi(x^n)$, we write $\mathcal{P}_\phi$ for the

PML estimate $\mathcal{P}_\varphi(x^n)$. According to Theorem 3, for any parameters $\varepsilon > 0$ and $\delta \in (0,1)$, if there exists an estimator $\hat{f}$ such that

$$\Pr_{X^n \sim p}\left(\left|\hat{f}(X^n) - f(p)\right| > \varepsilon\right) < \delta,\ \forall p \in \mathcal{P},$$

there is an estimator $\hat{f}_\varphi$ over profiles satisfying

$$\Pr_{X^n \sim p}\left(\left|\hat{f}_\varphi(\varphi(X^n)) - f(p)\right| > \varepsilon\right) < \delta,\ \forall p \in \mathcal{P}.$$

For an arbitrary length-n profile ϕ that satisfies $\Pr_{\Phi^n \sim p}(\Phi^n = \phi) \geq 2\delta$, these error bounds yield $\Pr(|\hat{f}_\varphi(\phi) - f(p)| > \varepsilon) < \frac{1}{2}$, and since $\Pr_{\Phi^n \sim \mathcal{P}_\phi}(\Phi^n = \phi) \geq \Pr_{\Phi^n \sim p}(\Phi^n = \phi) \geq 2\delta$ by the definition of PML (as we take the distribution that maximizes the probability),

$$\Pr\left(\left|\hat{f}_\varphi(\phi) - f(\mathcal{P}_\phi)\right| > \varepsilon\right) < \frac{1}{2}.$$

By the union bound and triangle inequality,

$$\Pr\left(\left|f(p) - f(\mathcal{P}_\phi)\right| > 2\varepsilon\right) < 1 \iff \left|f(p) - f(\mathcal{P}_\phi)\right| \leq 2\varepsilon \ \text{ surely.}$$

Furthermore, by Lemma 8, with probability at least $1 - 6/\sqrt{n}$, the total probability of length-n profiles ϕ satisfying $\Pr_{\Phi^n \sim p}(\Phi^n = \phi) < 2\delta$ is at most

$$2\delta \cdot C_{\frac{6}{\sqrt{n}},n}(p) + \frac{6}{\sqrt{n}} \leq 2\delta \cdot n^{8\mathcal{D}_n + 20\log n} + \frac{6}{\sqrt{n}},$$

which basically upperly bounds the probability that $|f(p) - f(\mathcal{P}_{\Phi^n})| > 2\varepsilon$. Next we will assume that there exists an estimator $\hat{f}$ satisfying $\Pr_{X^m \sim p}(|\hat{f}(X^m) - f(p)| > \varepsilon) < \delta,\ \forall p \in \mathcal{P}$. By Lemma 7, if $\delta \leq 1/10$, we can construct another estimator $\hat{f}'$ that takes a sample of size $n = \frac{4m}{\log \frac{1}{2\delta}} \log \frac{1}{\delta'}$ (n

is assumed to be an integer here) and achieves a higher-confidence guarantee

$$\Pr_{X^n\sim p}\left(\left|\hat{f}'(X^n)-f(p)\right|>\varepsilon\right)<\delta',\ \forall p\in\mathcal{P}.$$

Then by the above reasoning, with probability at least $1-6/\sqrt{n}$,

$$\begin{aligned}\Pr_{\Phi^n\sim p}\left(|f(p)-f(\mathcal{P}_{\Phi^n})|>2\varepsilon\right)&\le 2\delta'\cdot n^{8\mathcal{D}_n+20\log n}+\frac{6}{\sqrt{n}}\\&=2\exp\left(-\frac{n}{4m}\log\frac{1}{2\delta}+(8\mathcal{D}_n+20\log n)\log n\right)+\frac{6}{\sqrt{n}}.\end{aligned}$$

For the first term on the right hand side to vanish as quickly as $1/\sqrt{n}$, it suffices to have

$$\frac{n}{4m}\log\frac{1}{2\delta}\ge 20\cdot\mathcal{D}_n\log n \text{ and } \frac{n}{4m}\log\frac{1}{2\delta}\ge 40\cdot\log^2 n.$$

Simplifying the expressions and applying the union bound yield that $|f(p)-f(\mathcal{P}_{\Phi^n})|\le 2\varepsilon$ with probability at least $1-1/\sqrt{n}$, given both

$$\frac{n}{\mathcal{D}_n}\gtrsim\frac{m}{\log\frac{1}{\delta}} \text{ and } n\ge 8m. \qquad \square$$

5.B.5 Experiments

Prior works such as [HO19a, PJW17] have experimentally demonstrated the efficiency of PML on estimating several classical properties, including the Shannon and Rényi entropy, support size, and ℓ_1 distance to the uniform distribution. Our result further extends and establishes the efficiency of PML for numerous symmetric properties that are under-explored. Given the broadness of this property class, the potential applications are countless.

Consider a variant of Shannon entropy, $f(p):=\sum_x p_x\log^2 p_x$, that mildly puts more emphasis on small probabilities. As the property is relatively new and non-Lipschitz, prior works

and approaches do not easily yield a satisfiable learning guarantee. Our result hence comes into play, because f is symmetric, which suffices for Theorem 4 to take effect. Below, we will estimate this property by an n-sample PML plug-in, and compare its performance to two estimators: the n-sample empirical estimator that evaluates the entropy of the empirical distribution, serving as a baseline, and the $10n$-sample empirical estimator whose sample size is larger than others by *an order of magnitude*.

We considered six natural distributions: uniform, Zipf(1/2), Zipf(2), Dirichlet(1)-drawn-, Dirichlet(2)-drawn-, and geometric, all having support size $k = 5{,}000$. The plots are presented in Figure 5.1, with both vertical and horizontal axes showing in *log-scale* (base 10). The sample size n ranges from 10^3 to 10^5, and every data point represents the average absolute error over 20 independent simulations.

Specifically, the geometric distribution has a success probability of $(k-1)/k$; the Zipf(1/2) and Zipf(2) distributions have probability $p_i \propto i^{-1/2}$ and $p_i \propto i^{-2}$ for $i \geq 1$, both being truncated at k and re-normalized; drawing a distribution from the Dirichlet(1) prior is equivalent to drawing one uniformly from the k-dimensional standard simplex.

As the experiments demonstrate, the PML plug-in estimator significantly improves over the empirical estimator (note that the axes are in log-scale) and is as good as an estimator having access to samples larger by order of magnitudes. There are multiple PML implementations and we have used the one by [HO19a] (Section 4 of that paper presents a list of PML computation algorithms). Code is included in the supplementary material. For instructions on how to use the code, please refer to the inline comments and Section 4.1 in the supplementary material of [HO19a].

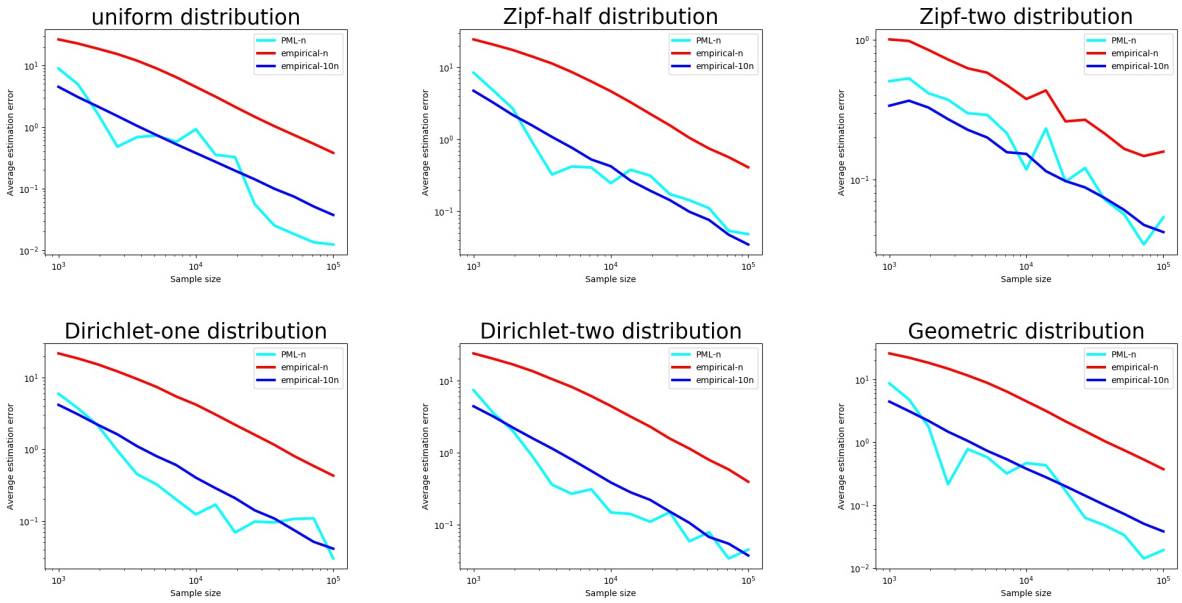


Figure 5.1: Inferring property f via the PML plug-in.

5.C Competitive Estimation of Distributions and Entropy

5.C.1 Theorem 2: Competitive Distribution Estimation

Estimating distributions from their samples is a statistical-inference cornerstone, and has numerous applications, ranging from biological studies [AIS+08] to language modeling [CG99]. A learning algorithm $\hat{p}$ in this setting is called a *distribution estimator*, which associates with every sequence x^n a distribution $\hat{p}(x^n) \in \Delta$. Given a sample $X^n \sim p$, we measure the performance of $\hat{p}$ in estimating distribution p by the Kullback-Leibler (KL) divergence $D(p \parallel \hat{p}(X^n))$.

Let $r_n(p, \hat{p}) := \min\{r : \Pr(D(p \parallel \hat{p}(X^n)) \le r) \ge 9/10\}$ be the *minimal KL error* $\hat{p}$ could achieve with probability at least $9/10$. Then, the *worst-case error* of estimator $\hat{p}$ over $\mathrm{P} \subseteq \Delta$ is $r_n(\mathrm{P}, \hat{p}) := \max_{p \in \mathrm{P}} r_n(p, \hat{p})$, and the lowest worst-case error for P, achieved by the optimal estimator, is the *minimax error* $r_n(\mathrm{P}) := \min_{\hat{p}'} r_n(\mathrm{P}, \hat{p}')$. The most widely studied distribution set P is simply $\Delta_{\mathcal{X}}$. With $\mathcal{X}$ being finite, it has become a classical result that $r_n(\Delta_{\mathcal{X}}) = \Theta(|\mathcal{X}|/n)$, which is achievable, up to constant factors, by an add-constant estimator [BS04, KOPS15].

Beyond minimax Despite being minimax optimal, the $|\mathcal{X}|/n$-result and the algorithm, are not satisfiable from a practical point of view. The reason is that the formulation puts much of its emphasis on the worst-case performance, and ignores the intrinsic simplicity of p in a pessimistic fashion. Hence, the desire to design more efficient estimators for practical distributions, like power-law, or Poisson, has led to algorithms that possess adaptive estimation guarantees.

Concretely, the minimax formulation has two modifiable components – the collection P and the error function D. A common approach to specifying P is adding structural assumptions, such as monotonicity, m-modality, and log-concavity, which, in many cases, makes algorithm refinement possible by leveraging structural simplicity. An orthogonal approach to encouraging adaptability without imposing structures is to replace absolute error by relative error, which we illustrate below.

Competitive estimation Without strong prior knowledge on the underlying distribution, a reasonable estimator should *naturally* assign the same probability to symbols appearing an equal number of times. *Competitive estimation* calls for finding a universally near-optimal estimator that learns *every* distribution as well as the best natural estimator that knows the true distribution.

Denote by $\mathcal{N}$ the collection of all natural estimators. For any distribution $p \in \Delta$ and sample $X^n \sim p$, a given estimator $\hat{p}$ incurs, with respect to the best natural estimator knowing p, an instance-by-instance *relative KL error* of

$$\mathrm{D}_{\text{nat}}(p \parallel \hat{p}(X^n)) := D(p \parallel \hat{p}(X^n)) - \min_{\hat{q} \in \mathcal{N}} D(p \parallel \hat{q}(X^n)).$$

Analogous to the minimax formulation, we denote by

$$r_n^{\text{nat}}(p, \hat{p}) := \min\{r : \Pr(\mathrm{D}_{\text{nat}}(p \parallel \hat{p}(X^n)) \le r) \ge 9/10\}$$

the *minimal relative error* $\hat{p}$ achieves with probability at least $9/10$, by $r_n^{\text{nat}}(\mathrm{P}, \hat{p})$ the *worst-case relative error* of $\hat{p}$ over $\mathrm{P} \subseteq \Delta$, and by $r_n^{\text{nat}}(\mathrm{P})$ the *minimax relative error*.

Old and new results Initiating the competitive formulation, [OS15] show that a simple variant of the well-known Good-Turing estimator achieves $r_n^{\text{nat}}(\Delta) \lesssim 1/n^{1/3}$, and a more involved estimator in [AJOS13b] attains the optimal $r_n^{\text{nat}}(\Delta) \simeq 1/\sqrt{n}$. For a fully adaptive guarantee, [HO19b] further refine the bound and design an estimator $\hat{p}^\star$ achieving $r_n^{\text{nat}}(p,\hat{p}^\star) \lesssim \mathbb{E}_{\mathcal{D}_n\sim p}[\mathcal{D}_n/n] \lesssim r_n^{\text{nat}}(\Delta)$, for every $p \in \Delta$, but provide no lower bounds.

In this work, we completely characterize $r_n^{\text{nat}}(p,\cdot)$ with essentially matching lower and upper bounds. Surprisingly, we show that for nearly every sample size n, the quantity behaves like $\mathcal{H}_n(p)/n$.

Theorem 2 (Optimal competitive error)**.** *There is a near-linear-time computable estimator $\hat{p}^\star$, such that for any distribution p and n,*

$$r_n^{nat}(p,\hat{p}^\star) \lesssim \frac{\mathcal{H}_n(p)}{n}.$$

where $\hat{p}^\star$ is the near linear-time computable estimator in [HO19b] mentioned above. On the other hand, for any $H \in [0,\sqrt{n})$,

$$\min_{\hat{p}} \max_{p:\mathcal{H}_n(p)\lesssim H} r_n^{nat}(p,\hat{p}) \gtrsim \frac{H}{n}.$$

5.C.2 Proof of Theorem 2

Proof. The upper bound follows by the main result of [HO19b] and Theorem 1 asserting the entropy-dimension equivalence. To establish the lower bound, denote $s := (H/\log n)^{1/2}$, $I := \{s, s+1, \ldots, 2s\}$, and $P := \cup_{i\in I} P_i := \cup_{i\in I} U_i/n$ where

$$U := \bigcup_{i\in I} U_i := \bigcup_{i\in I} \{i^2\log^2 n, i^2\log^2 n + 1, \ldots, i^2\log^2 n + i\log n\},$$

where $H \lesssim \sqrt{n/\log n}$ for the total to be at most n. Let $A \cdot \{B\}$ denote the length-A constant sequence of value B. Let C be the set of distributions in the form of

$$p := L \cdot \left\{\frac{1}{n^2}\right\} \bigcup \left(\bigcup_i (i \log n) \cdot \left\{ q_i \text{ or } q_i' : nq_i = i^2 \log^2 n, nq_i' = i^2 \log^2 n + i \log n \right\} \right).$$

where the probability values are sorted according to the ordering they appear above, L is a proper variable that makes the probabilities sum to 1, and the range of support of distribution p is irrelevant for our purpose and hence unspecified. Equip a uniform prior over C (equivalently, construct a random distribution). We have several claims in order:

- For any $i \in I$ and $\mu \in U_i$, by the construction and independence,

$$\begin{aligned}
\Pr(\varphi_\mu = 1 | q_i \text{ is chosen}) &\approx (i \log n) \cdot \left(\Pr(\mathrm{Poi}(nq_i) = \mu) \cdot (\Pr(\mathrm{Poi}(nq_i) \neq \mu))^{i \log n - 1} \right) \\
&\approx (i \log n) \cdot \left(\frac{1}{\sqrt{nq_i}} \cdot \left(1 - \frac{1}{\sqrt{nq_i}}\right)^{i \log n - 1} \right) \\
&\geq \Omega(1).
\end{aligned}$$

  Similarly, we have $\Pr(\varphi_\mu = 1 | q_i' \text{ is chosen}) \geq \Omega(1)$. Hence,

$$\Pr(\varphi_\mu = 1) \geq \Omega(1).$$

- For any $i \in I$ and $\mu \in U_i$, by Bayes' rule,

$$\Pr(q_i \text{ is chosen} | \varphi_\mu = 1) = \frac{\Pr(\varphi_\mu = 1 | q_i \text{ is chosen}) \cdot 0.5}{\Pr(\varphi_\mu = 1)} \geq \Omega(1).$$

  Similarly, we have $\Pr(q_i' \text{ is chosen} | \varphi_\mu = 1) \geq \Omega(1)$.

- For any $i \in I$ and $\mu \in U_i$, the value of M_μ, the total probability of symbols appearing μ times, is q_i if $\varphi_\mu = 1$ and q_i is chosen; and is q_i' if $\varphi_\mu = 1$ and q_i is chosen. Any estimator E_μ will

incur an expected absolute error of $\Omega(i(\log n)/n)$ in estimating M_μ given $\varphi_\mu = 1$.

- Note that for any $\alpha \in [0,1]$ and $x, y > 0$,

$$\alpha(y-z)^2 + (1-\alpha)(z-x)^2 \geq \alpha(1-\alpha)(x-y)^2.$$

- Therefore, the expected squared Hellinger distance $\mathbb{H}^2(\cdot,\cdot)$ of any estimator E_μ in estimating $(M_\mu)_{\mu \geq 0}$ satisfies, by the linearity of expectation,

$$\begin{aligned}
\frac{1}{2}\sum_{\mu \geq 0} \mathbb{E}\left(\sqrt{E_\mu} - \sqrt{M_\mu}\right)^2 &\geq \frac{1}{2}\sum_{i \in I}\sum_{\mu \in U_i} \mathbb{E}\left[\left(\sqrt{E_\mu} - \sqrt{M_\mu}\right)^2 \middle| \varphi_\mu = 1\right] \Pr(\varphi_\mu = 1) \\
&= \frac{1}{2}\sum_{i \in I}\sum_{\mu \in U_i} \mathbb{E}\left[\left(\frac{E_\mu - M_\mu}{\sqrt{E_\mu} + \sqrt{M_\mu}}\right)^2 \middle| \varphi_\mu = 1\right] \Pr(\varphi_\mu = 1) \\
&\geq \sum_{i \in I} (i \log n) \cdot \Omega\left(\frac{(i\log n)/n}{\sqrt{i^2(\log^2 n)/n}}\right)^2 \\
&\geq s \cdot \Omega\left(\frac{s \log n}{n}\right) \\
&= \Omega\left(\frac{H}{n}\right).
\end{aligned}$$

- Consequently, by the inequality $D(P \parallel Q) \geq 2\mathbb{H}^2(P,Q)$,

$$\mathbb{E}\left[D(E \parallel M)\right] \geq \mathbb{E}\left[2\mathbb{H}^2(E,M)\right] \geq \Omega\left(\frac{H}{n}\right).$$

- Finally, combining Theorem 1, 9 and 1 yields that, with high probability,

$$\mathcal{H}_n(p) \simeq \mathcal{D}_n \simeq E_n(p) \simeq \mathcal{H}_n^s(p) = \sum\nolimits_{j \geq 1} \min\left\{p_{I_j}, j \cdot \log n\right\},$$

which, by our definition, is at most $O(\log n + s(s \log n)) = O(\log n + H)$. □

5.C.3 Extension: Competitive Entropy Estimation

Recall that a distribution estimator is *natural* if it assigns the same probability to symbols of equal multiplicity, and a property estimator is *plug-in* if it first finds an estimate of the distribution and then evaluates the property at this estimate. As an off-the-shelf method, the plug-in approach is widely used in estimating distribution properties.

As we mentioned in Appendix 5.B.3, to estimate a symmetric property, an accurate estimate of the probability multiset of the underlying distribution suffices. Intuitively, it should be easier in terms of statistical efficiency to recover just the probability mutiset than to learn the entire distribution. For example, over distribution collection $\Delta_{\mathcal{X}}$, the PML plug-in estimator is minimax optimal for learning entropy, while the empirical distribution, being minimax optimal for distribution estimation, is suboptimal as a plug-in entropy estimator.

However, the analysis and computation (though efficient) of such multiset-based estimation methods are often involved [VV11a, VV13, VV16, HJW18, CSS19b, ADOS17, HO19a]. For this reason, plug-in estimators that first estimate the true distribution are still popular in practice, and often, the distribution component is natural.

For example, several notable and widely used entropy estimators are *natural plug-in*, including the empirical estimator that simply uses the empirical distribution, James-Stein shrinkage [HS09] that shrinks the distribution estimate towards uniform, and Dirichlet-smoothed [SG96] that imposes a Dirichlet prior over $\Delta_{\mathcal{X}}$.

The logic behind these estimators is simple – if two distributions (e.g., the true distribution and our estimate) are close, the same is expected for their entropy values. The next theorem shows that for *every* distribution and among all plug-in entropy estimators, the distribution estimator in [HO19b] is as good as the one that performs best in estimating the actual distribution.

Denote by $\mathcal{N}$ the collection of all natural estimators, and write $|H(p) - H(q)|$ as $\ell_H(p, q)$.

Theorem 13 (Competitive entropy estimation)**.** *For any distribution p, sample $X^n \sim p$, and the*

respective best natural estimator $\hat{p}^{\mathcal{N}}_{X^n} := \arg\min_{\hat{p}\in\mathcal{N}} D(p \parallel \hat{p}_{X^n})$*, with probability at least* $1-1/n$*,*

$$\ell_H(p,\hat{p}^{\star}_{X^n}) - \ell_H(p,\hat{p}^{\mathcal{N}}_{X^n}) \le \tilde{O}\left(\sqrt{\frac{\mathcal{H}_n(p)}{n}}\right).$$

Proof. Given any natural estimator and a sample $X^n \sim p$, we denote by q the distribution estimate. The entropy of q differs from the true entropy by

$$\begin{aligned}
H(q) - H(p) &= -\sum_x q_x \log q_x + \sum_x p_x \log p_x \\
&= \sum_x p_x \log p_x - \sum_x p_x \log q_x + \sum_x p_x \log q_x - \sum_x q_x \log q_x \\
&= \sum_x p_x \log \frac{p_x}{q_x} + \sum_x (p_x - q_x) \log q_x \\
&= D(p \parallel q) + \sum_x (p_x - q_x) \log q_x.
\end{aligned}$$

Denote by $P_\mu(X^n)$ and $Q_\mu(X^n)$ the total probability that distributions p and q assign to symbols with multiplicity μ. Since q is induced by a natural estimator, we also write $q_\mu(X^n)$ for the probability that q assigns to *each* symbol with multiplicity μ in X^n. Recall that prevalence $\varphi_\mu(X^n)$ denotes the number of symbols with multiplicity μ in X^n. Therefore, $Q_\mu(X^n) = \varphi_\mu(X^n) \cdot q_\mu(X^n)$.

Henceforth, whenever it is clear from the context, we suppress X^n in related expressions. Then, the second term on the right-hand side satisfies

$$\begin{aligned}
\sum_x (p_x - q_x) \log q_x &= \sum_x \Big(\sum_\mu \mathbb{1}_{\mu_x=\mu} \cdot p_x - \sum_\mu \mathbb{1}_{\mu_x=\mu} \cdot q_\mu\Big) \log\Big(\sum_\mu \mathbb{1}_{N_x=\mu} \cdot q_\mu\Big) \\
&= \sum_x \sum_\mu \mathbb{1}_{\mu_x=\mu} \cdot (p_x - q_\mu) \log q_\mu \\
&= \sum_\mu \Big(\sum_x \mathbb{1}_{\mu_x=\mu} \cdot p_x - \sum_x \mathbb{1}_{\mu_x=\mu} \cdot q_\mu\Big) \log q_\mu \\
&= \sum_\mu \big(P_\mu - Q_\mu\big) \log q_\mu.
\end{aligned}$$

Let $q_{\min}$ be the smallest nonzero probability of q. By the triangle inequality and Pinsker's

inequality,

$$\begin{aligned}\left|\sum_{\mu}\big(P_\mu - Q_\mu\big)\log q_\mu\right| &\le \sum_{\mu}\big|\big(P_\mu - Q_\mu\big)\log q_\mu\big| \\ &\le |\log q_{\min}|\sum_{\mu}\big|P_\mu - Q_\mu\big| \\ &\le |\log q_{\min}|\sqrt{2D(P\parallel Q)}.\end{aligned}$$

For simplicity, suppress the subscript X^n from all estimators, e.g., write $\hat{p}^{\mathcal{N}} := \hat{p}^{\mathcal{N}}_{X^n}$. Now we show that if a symbol x has multiplicity μ, the estimator $\hat{p}^{\mathcal{N}}$ will assign a probability mass of P_μ/φ_μ. In other words, $\hat{P}^{\mathcal{N}}_\mu = P_\mu$ since $p^{\mathcal{N}} \in \mathcal{N}$. Indeed, the corresponding KL-divergence values differ by

$$\begin{aligned}\sum_x p_x \log\frac{p_x}{q_x} - \sum_x\sum_\mu \mathbb{1}_{\mu_x=\mu}\cdot p_x\log\frac{p_x}{P_\mu/\varphi_\mu} &= \sum_x p_x\log\frac{1}{q_x} - \sum_x\sum_\mu \mathbb{1}_{\mu_x=\mu}\cdot p_x\log\frac{\varphi_\mu}{P_\mu} \\ &= \sum_x\sum_\mu \mathbb{1}_{\mu_x=\mu}\cdot p_x\log\frac{P_\mu}{\varphi_\mu q_\mu} \\ &= \sum_\mu P_\mu\log\frac{P_\mu}{Q_\mu} = D(P\parallel Q)\ge 0.\end{aligned}$$

Then, the above equalities yield that,

$$H(\hat{p}^{\mathcal{N}}) - H(p) = D(p\parallel\hat{p}^{\mathcal{N}}) + \sum_\mu\left(P_\mu - \hat{P}^{\mathcal{N}}_\mu\right)\log p^{\mathcal{N}}_\mu = D(p\parallel\hat{p}^{\mathcal{N}}).$$

Next consider the other estimator $\hat{p}^\star$, which is also natural. Let $\mathcal{D}_n = \mathcal{D}_n$ be the profile dimension of X^n. By the results in [HO19b], estimator $\hat{p}^\star$ achieves a $\mathcal{D}_n/n$ excess loss, i.e.,

$$D(p\parallel\hat{p}^\star_{X^n}) - \min_{\hat{p}\in\mathcal{N}} D(p\parallel\hat{p}_{X^n}) = D(P\parallel\hat{P}^\star) \le \tilde{O}\left(\frac{\mathcal{D}_n}{n}\right),$$

for every p and $X^n\sim p$, with probability at least $1 - O(1/n)$. In addition, by its construction, the

minimum probability of $\hat{p}_{X^n}$ is at least $1/n^4$. Therefore, with probability at least $1 - O(1/n)$,

$$\left|\sum_x (p_x - \hat{p}^\star_x)\log \hat{p}^\star_x\right| = \left|\sum_\mu \left(P_\mu - \hat{P}^\star_\mu\right)\log \hat{p}^\star_\mu\right| \le |\log \hat{p}^\star_{\min}| \cdot \sqrt{2D(P \parallel \hat{P}^\star)} \le \tilde{O}\left(\sqrt{\frac{\mathcal{D}_n}{n}}\right).$$

Finally, the triangle inequality combines the above results and yields

$$\begin{aligned}
\ell_H(p,\hat{p}^\star) - \ell_H(p,\hat{p}^{\mathcal{N}}) &= |H(p) - H(\hat{p}^\star)| - |H(p) - H(\hat{p}^{\mathcal{N}})| \\
&= \left|D(p \parallel \hat{p}^\star_x) + \sum_x (p_x - \hat{p}^\star_x)\log \hat{p}^\star_x\right| - \left|\min_{\hat{p}\in\mathcal{N}} D(p \parallel \hat{p})\right| \\
&\le \left|D(p \parallel \hat{p}^\star_x) - \min_{\hat{p}\in\mathcal{N}} D(p \parallel \hat{p})\right| + \left|\sum_x (p_x - \hat{p}^\star_x)\log \hat{p}^\star_x\right| \\
&= D(P \parallel \hat{P}^\star_\mu) + \tilde{O}\left(\sqrt{\frac{\mathcal{D}_n}{n}}\right) \\
&\le \tilde{O}\left(\sqrt{\frac{\mathcal{D}_n}{n}}\right).
\end{aligned}$$

This together with Theorem 1 completes the proof. □

5.C.4 Experiments

The experiments in [HO19b] have demonstrated the efficiency of $\hat{p}^\star$, showing that the estimator frequently and uniformly outperforms an improved version of the well-known Good-Turing estimation scheme [OS15], for numerous distributions and parameter settings. Our results confirmed the optimality of estimator $p^\star$ from a theoretical point of view, and moves forward considerably our understanding of how well one can approach the performance of a genie having the full knowledge of the true distribution, but restricted to be natural as all human beings.

In the following, we do not repeat the experiments in [OS15] (see Section 2 of its supplementary), and instead, investigate a novel and highly related task – employing $\hat{p}^\star$ as a plug-in estimator for Shannon entropy. By Theorem 13 and its proof, we already see that the

resulting plug-in estimator $H \circ \hat{p}^\star$ is as good as any plug-in estimator with a natural distribution component, and how well it performs, to a certain extent, depends on how well it approximates the true distribution under the KL divergence. But is this plug-in estimator still competitive when compared to estimators having observed samples of much larger sizes, or to the state-of-the-art estimators that are designed just for entropy estimation? The following experiments answered this question in the affirmative.

Below we demonstrate the efficiency of $\hat{p}^\star$ when used as a plug-in entropy estimator. We will compare its performance with a size-n sample to three estimators: the n-sample *empirical* estimator that evaluates the entropy of the empirical distribution, the $n \log n$-sample empirical estimator that has access to much more information, and a state-of-the-art entropy estimator in [WY16] based on minimax polynomial approximations (which we refer to as WY). Shown by the experiments in [WY16], under numerous settings, the WY estimator frequently outperformed several classical estimators and other minimax estimators such as [VV11a, VV13, JVHW15]. Hence, we maintain simplicity and do not compare our approach to the latter ones.

We considered six natural distributions: uniform, two-steps-, Zipf(1/2), binomial, geometric, and Dirichlet(1)-drawn-, all having support size $k = 5{,}000$. The plots are presented in Figure 5.2, with both vertical and horizontal axes showing in *log-scale* (base 10). The sample size n ranges from 10^3 to 10^5, and every data point represents the average absolute error over 20 independent simulations. We refer to the plug-in estimator using $\hat{p}^\star$ as *HO*.

Specifically, 10% probability values of the two-steps distribution $\propto 9/k$, and the rest $\propto 1/k$; the binomial and geometric distributions have success probabilities of $10/k$ and $(k-1)/k$, respectively; the Zipf(1/2) distribution has probability $p_i \propto i^{-1/2}$ for $i \geq 1$, and is truncated at k and re-normalized.

We see that the performance of the WY estimator and our plug-in approach are essentially the same. In particular, for Dirichlet(1)-drawn-, WY is better, but for binomial, WY is worse; for all other cases, the two error curves basically follow the same trend and lie in the same region.

This is somewhat surprising since intuitively, $\hat{p}^\star$ is a distribution estimator and its design has no consideration about entropy estimation, while WY is geared towards this task. On the other hand, the performance of the induced plug-in estimator should be both efficient and competitive, as guaranteed by Theorem 13.

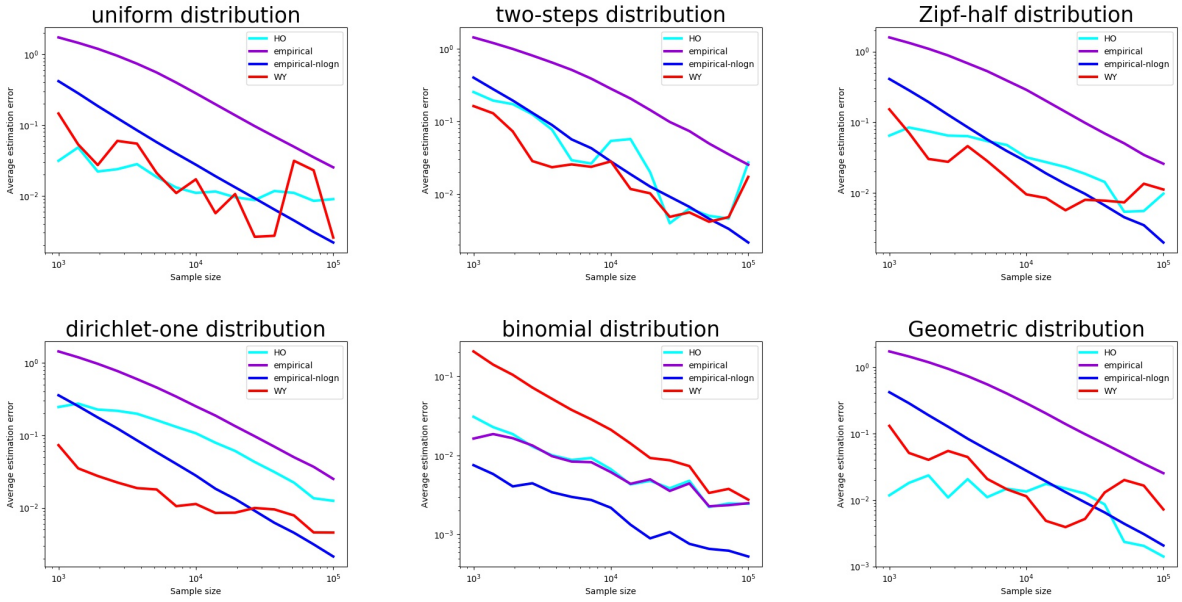


Figure 5.2: Competitive entropy estimation

5.D Optimal Characterization for Structured Families

Following the previous discussions, we will derive nearly tight bounds on $\mathcal{H}_n(p)$ for three important structured families – log-concave, power-law, and histogram. These bounds clearly demonstrate the power of profile entropy in charactering natural shape constraints.

For the subsections below, we adopt the convention of specifying structured distributions over $\mathcal{X} = \mathbb{Z}$.

5.D.1 Theorem 6: Log-Concave Family

The log-concave family encompasses a broad range of discrete distributions, such as Poisson, hyper-Poisson, Poisson binomial, binomial, negative binomial, and geometric, and hyper-geometric, with broad applications to statistics [SW14], computer science [LV07], economics [An97], and geometry [Sta89].

Formally, a distribution $p \in \Delta_{\mathbb{Z}}$ is *log-concave* if p has a contiguous support and $p_x^2 \geq p_{x-1} \cdot p_{x+1}$ for all $x \in \mathbb{Z}$. The next result bounds the profile entropy of this family, and is *tight* up to logarithmic factors. For simplicity, henceforth we write $a \wedge b$ for $\min\{a,b\}$ (and $\vee$ for max), and slightly abuse the notation and write $a \simeq b$ for $a+1 = \tilde{\Theta}(b+1)$, which does not change the nature of the results.

Theorem 6. *Let $\mathcal{L}_\sigma \subseteq \Delta_{\mathbb{Z}}$ denote the collection of log-concave distributions with variance σ^2. Then,*

$$\max_{p \in L_\sigma} \mathcal{H}_n(p) \simeq \sigma \wedge \frac{n}{\sigma}.$$

In particular, if we discretize a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma^2)$ by rounding it to the nearest integer, the distribution of the resulting variable achieves the maximum, up to logarithmic factors. Moreover, such a discretization procedure preserves log-concavity for any continuous distribution over $\mathbb{R}$.

A similar bound holds for t-mixtures of log-concave distributions. More concretely,

Theorem 14. *For any t-mixture $p \in \Delta_{\mathbb{Z}}$ of log-concave distributions with variances $\sigma_i^2, 1 \leq i \leq t$,*

$$\mathcal{H}_n(p) \lesssim \left(\sum_i \sigma_i \right) \wedge \max_i \left\{ \frac{n}{\sigma_i} \right\},$$

where the right-hand side is assumed to be at least t since otherwise $\mathcal{H}_n(p) \lesssim t$, and in practice, t is often a small quantity, e.g. a constant.

5.D.2 Proof of Theorem 6 and 14

We start by showing the $\mathcal{H}_n(p) \gtrsim \sigma \wedge n/\sigma$ lower bound. A requirement is that p must be a discrete log-concave distribution. We show that one can take p as a discretized Gaussian $\mathcal{N}(\mu, \sigma^2)$. In addition, the discretization procedure works for any continuous distribution and preserves log-concavity and essentially also the variance. We will start by introducing the discretization procedure.

Proof. Log-concavity is a generic structure exhibited by numerous classical distributions, both those discrete (introduced above) and continuous ones, such as Gaussian, exponential, uniform, logistic, and Laplace distributions. Below, we present a discretization procedure that preserves distribution shapes such as monotonicity, modality, and log-concavity. Applying this procedure to a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ yields the lower bound in Theorem 6.

Let X be a continuous random variable with density function $f(x)$. For any $x \in \mathbb{R}$, denote by $\lceil x \rfloor$ the closest integer z such that $x \in (z - 1/2, z + 1/2]$. The distribution of $\lceil X \rfloor$ is over $\mathbb{Z}$ and satisfies

$$p(z) := \int_{z-\frac{1}{2}}^{z+\frac{1}{2}} f(x)dx, \ \forall z \in \mathbb{Z}.$$

We refer to the random variable $\lceil X \rfloor$ as the discretized version of X.

Shape Preservation By the definition of $\lceil x \rfloor$, one can readily verify that the above procedure preserves several important shape characteristics of distributions, such as monotonicity, modality, and k-modality (possibly yields a smaller k). The following theorem further covers log-concavity.

Lemma 9. *For any continuous random variable X over $\mathbb{R}$ with a log-concave density f, the distribution $p \in \Delta_{\mathbb{Z}}$ associated with $\lceil X \rfloor$ is also log-concave.*

To show this, we need the following basic lemma about concave functions.

Lemma 10. *If f is a real concave distribution, for any real numbers x_1, x_2, y_1, and y_2 satisfying $x_1 \le x_2$, $y_1 \le y_2$, $x_1 < y_1$, and $x_2 < y_2$,*

$$\frac{f(y_1) - f(x_1)}{y_1 - x_1} \ge \frac{f(y_2) - f(x_2)}{y_2 - x_2}.$$

By the above lemma, for any $x, y \in \mathbb{R}$ such that $|x-y| \le 1$, and any function f that is log-concave, $\log f(x+1) - \log f(x) \le \log f(y) - \log f(y-1) \iff f(x+1)f(y-1) \le f(x)f(y)$.

Proof of Lemma 9. By definition, distribution p is log-concave if p has a consecutive support and $p(z)^2 \ge p(z+1)p(z-1), \forall z$. The first condition holds for $\lceil X \rfloor$ since X is has a continuous support on $\mathbb{R}$, and $p(z)$ is positive as long as $f(x) > 0$ for a non-empty sub-interval of $(z-1/2, z+1/2]$.

Below we show that p also satisfies the second condition. Specifically, for any $z \in \mathbb{Z}$,

$$\begin{aligned}
p(z-1)p(z+1) &= \left(\int_{z-\frac{3}{2}}^{z-\frac{1}{2}} f(x)dx\right)\left(\int_{z+\frac{1}{2}}^{z+\frac{3}{2}} f(x)dx\right) \\
&= \left(\int_{z-\frac{1}{2}}^{z+\frac{1}{2}} f(x-1)dx\right)\left(\int_{z-\frac{1}{2}}^{z+\frac{1}{2}} f(x+1)dx\right) \\
&= \int_{z-\frac{1}{2}}^{z+\frac{1}{2}}\int_{z-\frac{1}{2}}^{z+\frac{1}{2}} f(x-1)f(y+1)dxdy \\
&\le \int_{z-\frac{1}{2}}^{z+\frac{1}{2}}\int_{z-\frac{1}{2}}^{z+\frac{1}{2}} f(x)f(y)dxdy \\
&= \left(\int_{z-\frac{1}{2}}^{z+\frac{1}{2}} f(x)dx\right)^2 \\
&= p(z)^2,
\end{aligned}$$

where the inequality follows by Lemma 10 and its implication. □

Moment preservation Denote by p the distribution of $\lceil X \rfloor$ for $X \sim f$. Let μ and σ^2 be the mean and variance of density f, given that they exist. The theorem below shows that distribution p has, within small additive absolute constants, a mean of μ and variance of $\Theta(\sigma^2)$.

Lemma 11. *Under the aforementioned conditions, the mean of* $\lceil X \rfloor$ *satisfies*

$$\mathbb{E}\lceil X \rfloor = \mu \pm \frac{1}{2},$$

and the variance of $\lceil X \rfloor$ *satisfies*

$$(\sigma - 1)^2 \leq \mathbb{E}(\lceil X \rfloor - \mathbb{E}\lceil X \rfloor)^2 \leq (\sigma + 1)^2.$$

Proof of Lemma 11. First consider the mean value of $\lceil X \rfloor$ for $X \sim f$. We have

$$\mathbb{E}\lceil X \rfloor = \mathbb{E}[\lceil X \rfloor - X] + \mathbb{E}[X] = \mu \pm \frac{1}{2}.$$

Consider the variance of $\lceil X \rfloor$ and apply inequality $(a+b)^2 \leq a^2(1+1/t) + b^2(1+t), \forall t > 0$.

$$\begin{aligned}
\mathbb{E}(\lceil X \rfloor - \mathbb{E}\lceil X \rfloor)^2 &= \int_{-\infty}^{\infty} (\lceil x \rfloor - \mathbb{E}\lceil X \rfloor)^2 \cdot f(x)dx \\
&= \int_{-\infty}^{\infty} (\lceil x \rfloor - x + (x - \mathbb{E}X) + \mathbb{E}X - \mathbb{E}\lceil X \rfloor)^2 \cdot f(x)dx \\
&\leq \int_{-\infty}^{\infty} \left((\lceil x \rfloor - x + \mathbb{E}X - \mathbb{E}\lceil X \rfloor)^2 \left(1 + \frac{1}{t}\right) + (x - \mathbb{E}X)^2 (1+t) \right) f(x)dx \\
&\leq \int_{-\infty}^{\infty} \left(\left(1 + \frac{1}{t}\right) + (x - \mathbb{E}X)^2 (1+t) \right) f(x)dx \\
&= 1 + \frac{1}{t} + t\sigma^2 + \sigma^2 \\
&= (\sigma + 1)^2.
\end{aligned}$$

By a different inequality, $(a+b)^2 \geq a^2(1-1/t) + b^2(1-t), \forall t > 0$, we also have

$$\mathbb{E}(\lceil X \rfloor - \mathbb{E}\lceil X \rfloor)^2 \geq (\sigma - 1)^2. \qquad \square$$

By the above lemma, for almost any $\sigma \geq 1$, we can construct a discrete log-concave distribution of variance σ^2 if there is a continuous one with roughly the same variance.

Next, letting p_G denote the distribution of $\lceil X \rfloor$ for $X \sim \mathcal{N}(\mu, \sigma^2)$, we lower bound $\mathcal{H}_n^{\mathcal{S}}(p_G)$ (effectively, the profile entropy $\mathcal{H}_n(p_G)$). By definition, this discretized Gaussian, which we write as $\lceil \mathcal{N} \rfloor(\mu, \sigma^2)$, has a distribution in the form of

$$p_G(z) := \frac{1}{\sqrt{2\pi}\sigma} \int_{z-\frac{1}{2}}^{z+\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx, \ \forall z \in \mathbb{Z}.$$

Through the subsequent analysis, we show that

Lemma 12. *Under the aforementioned conditions,*

$$H_n^{\mathcal{S}}(p_G) \geq \Omega\left(\frac{1}{\log n}\right)\left(\sigma \wedge \frac{n}{\sigma}\right).$$

The lower bound in Theorem 6 follows by these inequalities.

Proof. At it is clear from the context, we write p instead of p_G. Recall that

$$H_n^{\mathcal{S}}(p) = \sum_{j \geq 1} \min\left\{p_{I_j}, j \cdot \log n\right\},$$

where p_{I_j} denotes the number of probabilities belonging to $I_j = ((j-1)^2, j^2] \cdot (\log n)/n$. Computing the quantity for part of the distribution can only reduce the value of $H_n^{\mathcal{S}}(p)$. Hence, we will focus on symbols in the $(\mu + 1, \infty) \cap \mathbb{Z}$ range, over which the probability mass function $p(z)$ is monotone.

We will further assume that $n/\log n \gg \sigma \gg \log n$, since otherwise the right-hand side of the inequality reduces to $O(1)$, and the result follows by $H_n^{\mathcal{S}}(p) \geq 1$ for all n and p. In addition, we focus on $j \gg 1$ in the following argument, as the contribution from $j = O(1)$ is relatively small.

Given these assumptions, we have

$$\begin{aligned} p(z) \in I_j &\iff \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z \pm 1/2 - \mu)^2}{2\sigma^2}\right) \in \left((j-1)^2 \frac{\log n}{n}, j^2 \frac{\log n}{n}\right] \\ &\iff z \pm 1/2 - \mu \in \sqrt{2}\sigma \left[\sqrt{c(\sigma,n) - 2\log j}, \sqrt{c(\sigma,n) - 2\log(j-1)}\right), \end{aligned}$$

where $c(\sigma,n) := \log\left(n/(\sqrt{2\pi}\sigma\log n)\right)$ and the interval is well-defined iff

$$c(\sigma,n) \geq 2\log j \iff \frac{n}{\sqrt{2\pi}\sigma\log n} \geq j^2 \iff \sqrt{\frac{n}{\sqrt{2\pi}\sigma\log n}} \geq j \impliedby \sqrt{\frac{n}{\sigma\log n}} \geq 2j.$$

For clarity, we divide our analysis into two cases: $\sqrt{n} \geq \sigma \gg \log n$ and $n/\log n \gg \sigma > \sqrt{n}$.

For the first case and $j \leq \sqrt{\sigma/\log n}/2 \leq \sqrt{n/(\sigma\log n)}/2$, the length L_j of the above interval, which equals to p_{I_j} up to an additive slack of 2, satisfies

$$\begin{aligned} \frac{L_j}{\sqrt{2}\sigma} &= \sqrt{c(\sigma,n) - 2\log(j-1)} - \sqrt{c(\sigma,n) - 2\log j} \\ &= \frac{2\log(j/(j-1))}{(c(\sigma,n) - 2\log(j-1)) + (c(\sigma,n) - 2\log j)} \\ &= \frac{\log(j/(j-1))}{\log\left(n/(\sqrt{2\pi}j(j-1)\sigma\log n)\right)} \\ &= \Omega\left(\frac{1}{\log n}\log\left(1 + \frac{1}{j-1}\right)\right) \\ &= \Omega\left(\frac{1}{j\log n}\right). \end{aligned}$$

Therefore, we have $L_j = \Omega(\sigma/(j\log n))$. Since $\sigma \gg \log n$ ensures $L_j \geq 3$ and $j \leq \sqrt{\sigma/\log n}/2$ is equivalent to $\sigma \geq 4j^2\log n$, the lower bound on L_j transforms into $p_{I_j} \geq \Omega(j)$. Hence in this case, $H_n^{\mathcal{S}}(p)$ admits the following bound

$$H_n^{\mathcal{S}}(p) = \sum_{j\geq 1} \min\left\{p_{I_j}, j \cdot \log n\right\} \geq \sum_{j=O(1)}^{\sqrt{\sigma/\log n}/2} \Omega(j) = \Omega\left(\frac{\sigma}{\log n}\right).$$

In the $n/\log n \gg \sigma > \sqrt{n}$ case, we have $\sqrt{\sigma/\log n} > \sqrt{n/(\sigma \log n)}$. Repeating the previous reasoning for $j \le \sqrt{n/(\sigma\log n)}/2$, we again obtain $L_j = \Omega\left(\sigma/(j\log n)\right)$ and $p_{I_j} \ge \Omega(j)$.

Therefore,

$$H_n^{\mathcal{S}}(p) = \sum_{j\ge 1} \min\left\{p_{I_j}, j\cdot\log n\right\} \ge \sum_{j=O(1)}^{\sqrt{n/(\sigma\log n)}/2} \Omega(j) = \Omega\left(\frac{n}{\sigma\log n}\right).$$

Finally, note that in the first case, $\min\{\sigma, n/\sigma\} = \sigma$, and in the second, $\min\{\sigma, n/\sigma\} = n/\sigma$.

Consolidating these results yields the desired lower bound

$$O(\log n)\cdot H_n^{\mathcal{S}}(p) \ge \sigma \wedge \frac{n}{\sigma}. \qquad \square$$

Next we proceed to the upper bound.

For any sample $X^n \sim p$, the profile dimension $\mathcal{D}(X^n)$ is at most the number of distinct symbols in the sample. It is well known that the tail probability of a log-concave distribution decays exponentially fast. Hence, the effective support size of p with respect to X^n is $\tilde{O}(\sigma+1)$, beyond which the tail probabilities can be as small as $1/n^3$ (the asymptotic notation hides logarithmic factors of n). Given this, even we sample from p for n times, the probability that we get only $\tilde{O}(\sigma+1)$ distinct symbols is at least $(1-1/n^3)^n \ge 1-1/n$. Therefore, we have $\mathcal{H}_n(p) \simeq \mathcal{D}(X^n) \lesssim \sigma+1$.

Now, we extend this argument to a t-mixture of log-concave distributions with variances $\sigma_i^2, i\in[t]$. For a length-n sample from this a distribution, the number of sample points from each mixture component is is at most n. Hence, with high probability, the number of distinct symbols in a length-n sample is at most $\sum\sigma_i + t$, up to logarithmic factors of n.

For the other part of the upper bound, we can assume that $\sigma \ge \sqrt{n}$ (otherwise we need to consider only the above case) and n is larger than some absolute constant. Then by a concentration inequality in [DKS16a], the maximum probability $p_{\max}$ of p belongs to $[1/(8\sigma), 1/\sigma]$. Hence,

the last index J for which $p_{I_J} \neq 0$ satisfies

$$(J-1)^2 \frac{\log n}{n} \leq \frac{1}{\sigma} \iff J \leq \sqrt{\frac{n}{\sigma \log n}} + 1.$$

Therefore, we have

$$\mathcal{H}_n^S(p) = \sum_{j \geq 1} \min\left\{p_{I_j}, j \cdot \log n\right\} \leq \log n + \sum_{j=1}^{\sqrt{n/(\sigma \log n)}+1} j \cdot \log n \leq O(\log n)\left(1 + \frac{n}{\sigma}\right).$$

Our upper bound is uniformly better than the $\min\{\sigma, (n^2/\sigma)^{1/3}\}$ bound in [HO19b], which is derived for $\mathcal{D}_n \sim p$. More importantly, we actually provide a complete characterization of the profile entropy value that is optimal up to logarithmic factors.

Next, we extend the n/σ bound to the mixture model. Write the mixture distribution as $p := \sum_i w_i \cdot p_i$, with w_i's being the mixing weights and p_i's being log-concave distributions with variances σ_i^2, respectively for $1 \leq i \leq t$. It is clear that $p_{\max}$ in this case is at most the maximum probability of some p_i, which at most $\max_i 1/\sigma_i$. The rest of the proof is the same as above. □

5.D.3 Theorem 7: Power-Law Family

Power-law Power-law is a ubiquitous structure appearing in many situations of scientific interest, ranging from natural phenomena such as the initial mass function of stars [Kro01], species and genera [HQD+10], rainfall [MR93], population dynamics [Tay61], and brain surface electric potential [MSODN09], to human-made circumstances such as the word frequencies in a text [Baa02], income rankings [DY01], company sizes [Axt01], and internet topology [FFF99].

Formally, a discrete distribution $p \in \Delta_{\mathbb{Z}}$ is a *power-law with power* $\alpha \geq 0$ if p has a support of $[k] := \{1, \ldots, k\}$ for some $k \in \mathbb{Z}^+ \cup \{\infty\}$ and $p_x \propto x^{-\alpha}$ for all $x \in [k]$. Note that if $\alpha \in [0,1]$, the distribution is well-defined for only finite k. The next result fully characterizes the profile entropy of power-laws over the entire ranges of α, n, and k.

Theorem 7. *Let $p \in \Delta_{[k]}$ be a power-law distribution with power α. Then,*

$$\mathcal{H}_n(p) \simeq \begin{cases} k & \text{if } \alpha > \frac{k^{1+\alpha}}{n} \vee 1 \text{ or } 1 \geq \alpha > \frac{k^2}{n}, \\ n^{\frac{1}{\alpha+1}} & \text{if } \frac{k^{1+\alpha}}{n} \geq \alpha > 1, \\ \left(\frac{n}{k^{1-\alpha}}\right)^{\frac{1}{1+\alpha}} & \text{if } \frac{k^2}{n} \wedge 1 \geq \alpha > \frac{k^{1-\alpha}}{n}, \\ \frac{n}{k^{1-\alpha}} - \frac{n}{k} & \text{if } \frac{k^{1-\alpha}}{n} \wedge 1 \geq \alpha \text{ and } \alpha \geq 2\log_k\left(7\sqrt{\frac{k}{n}}+1\right), \\ k \wedge \sqrt{\frac{n}{k^{1-\alpha}}} & \text{if } \frac{k^{1-\alpha}}{n} \wedge 1 \geq \alpha \text{ and } 2\log_k\left(7\sqrt{\frac{k}{n}}+1\right) > \alpha. \end{cases}$$

In particular, as $\alpha \to 0$, the bound degenerates to $k \wedge \sqrt{\frac{n}{k}}$, which is at most $n^{\frac{1}{3}}$.

Since a power-law sample profile is completely specified by α, k, and n, the above theorem directly applies to model parameter estimation. Specifically, we first compute $\mathcal{D}_n \sim p$, which is a simple function of the symbol counts. By Theorem 1, we can then use it to approximate $\mathcal{H}_n(p)$. Finally, we utilize the characterization theorem and find the parameter relations (testing might be necessary).

The theorem fully characterizes the profile entropy of power-laws and is significantly better than the basic $\{k, \sqrt{n\log n}\}$ bound for both $k \gg \sqrt{n}$ and $k \ll \sqrt{n}$. We can see how different parameter interplay with each other and leverage these relations in applications such as parameter estimation. In comparison, a result in [HO19b], when combined with our entropy-dimension equivalence theorem, yields only an $n^{1/(1+\alpha)}$ upper bound (and no lower bounds nor the right dependence on k), which is clearly suboptimal and provides no improvement over $\sqrt{n\log n}$ for $\alpha < 1$.

5.D.4 Proof of Theorem 7

Proof. For the ease of exposition, write the probability of symbol i assigned by distribution p as $p_i := c_\alpha^{-1} \cdot i^{-\alpha}$, where c_α is a normalizing constant that implicitly depends on k. Note that

$$\frac{k^{1-\alpha}}{1-\alpha} + \frac{\alpha}{1-\alpha} \geq 1 + \int_1^k x^{-\alpha} dx \geq c_\alpha = \sum_{i=1}^{k} i^{-\alpha} \geq \int_1^{k+1} x^{-\alpha} dx = \frac{(k+1)^{1-\alpha}}{1-\alpha} - \frac{1}{1-\alpha}.$$

By basic calculus and up to logarithmic factors, we can approximate the normalizing constant by

$$c_\alpha = \sum_{i=1}^{k} \frac{1}{i^\alpha} \simeq k^{1-\alpha} \vee 1,$$

Recall that the quantity of interest is essentially

$$H_n^{\mathcal{S}}(p) = \sum_{j \geq 1} \min\left\{ p_{I_j}, j \cdot \log n \right\}.$$

It will be convenient to denote $c := c(\alpha, k, n) := (c_\alpha \log n)/n \simeq (k^{1-\alpha} \vee 1)/n$. First, consider p_{I_j} for a sufficiently large j (i.e., $j \gg 1$) and note that

$$\begin{aligned} p_i \in I_j &\iff \frac{1}{c_\alpha i^\alpha} \in \left((j-1)^2 \frac{\log n}{n}, j^2 \frac{\log n}{n} \right] \\ &\iff i \in I_j' := \left[\left(j^2 c\right)^{-\frac{1}{\alpha}}, \left((j-1)^2 c\right)^{-\frac{1}{\alpha}} \right). \end{aligned}$$

Observe that the length L_j of interval I_j', which differs from the value of p_{I_j} by at most 2, is proportional to $(j-1)^{-2/\alpha} - j^{-2/\alpha}$, and hence is a decreasing function of j. Furthermore, each term $\min\{p_{I_j}, j \cdot \log n\} \approx \min\{L_j, j \cdot \log n\}$ is basically the minimum between this decreasing function and $j \log n$, an increasing function of j. This naturally calls for determining the value of

j at which the two functions are equal. Concretely,

$$\left((j-1)^2c\right)^{-\frac{1}{\alpha}} - \left(j^2c\right)^{-\frac{1}{\alpha}} = j\log n \Longrightarrow j \simeq J := \left(\frac{1}{\alpha^\alpha c}\right)^{\frac{1}{2+2\alpha}},$$

where J implicitly depends on α and n. In addition, since probability p_i vanishes if $i \notin [1,k]$, we need to consider only $\sqrt{1/(ck^\alpha)}+1 \le j \le \sqrt{1/c}$.

We can decompose the summation $H_n^{\mathcal{S}}(p)$ into two parts. The first part consists of indices $j \le J$,

$$H_{n,1}^{\mathcal{S}}(p) := \sum_{j=\sqrt{1/(ck^\alpha)}+1}^{J\wedge\sqrt{1/c}} \min\left\{p_{I_j}, j\cdot\log n\right\} \simeq \sum_{j=\sqrt{1/(ck^\alpha)}+1}^{J\wedge\sqrt{1/c}} j.$$

Correspondingly, the second part consists of indices $j \ge J$. For these indices j, we have $L_j \le j\cdot\log n$. Recall that I'_j specifies the range of i satisfying $p_i \in I_j$. Then the second part satisfies

$$H_{n,2}^{\mathcal{S}}(p) := \sum_{j=J\vee(\sqrt{1/(ck^\alpha)}+1)}^{\sqrt{1/c}} \min\left\{p_{I_j}, j\cdot\log n\right\} \simeq \sum_{j=J\vee(\sqrt{1/(ck^\alpha)}+1)}^{\sqrt{1/c}} L_j,$$

where the inequality follows by the fact that the intervals I'_j are consecutive. In addition, note that the left end point of I'_j equals $(J^2c)^{-\frac{1}{\alpha}} = (\alpha/c)^{\frac{1}{1+\alpha}}$.

The rest of the proof follows by dividing the analysis into several cases according to whether $\alpha > 1$ and the relative magnitude of J, $\sqrt{1/c}$, and $(\sqrt{1/(ck^\alpha)}+1)$.

For a concrete example, if $\alpha > 1$, then our approximation of c_α becomes $c_\alpha \simeq 1$, hence $c \simeq 1/n$, and it is also clear that $J = 1/(\alpha^\alpha c)^{\frac{1}{2\alpha+2}} \le \sqrt{1/c}$. Therefore,

$$H_{n,1}^{\mathcal{S}}(p) \simeq \sum_{j=\sqrt{1/(ck^\alpha)}+1}^{J} j.$$

Now, consider the relation between J and $\sqrt{1/(ck^\alpha)}$. By the continuity of profile entropy, we can treat c as $1/n$. If $\alpha \ge k^{1+\alpha}/n$, then $J \le \sqrt{1/(ck^\alpha)}$ and our upper bound for $H_{n,1}^{\mathcal{S}}(p)$ vanishes.

The quantity of interest hence becomes $H^{\mathcal{S}}_{n,1}(p)$, which equals to

$$H^{\mathcal{S}}_n(p) = H^{\mathcal{S}}_{n,2}(p) \simeq \sum_{j=\sqrt{1/(ck^\alpha)}+1}^{\sqrt{1/c}} L_j = k.$$

On the other hand, if $\alpha < k^{1+\alpha}/n$, then $J \geq \sqrt{1/(ck^\alpha)}+1$ and $H^{\mathcal{S}}_{n,1}(p)$ satisfies

$$H^{\mathcal{S}}_{n,1}(p) \simeq \sum_{j=\sqrt{1/(ck^\alpha)}+1}^{J} j \leq J^2 \simeq \left(\frac{n}{\alpha^\alpha}\right)^{\frac{1}{\alpha+1}}.$$

Our approximation of $H^{\mathcal{S}}_{n,2}(p)$ reduces to

$$H^{\mathcal{S}}_{n,2}(p) \simeq \sum_{j=J}^{\sqrt{1/c}} L_j \approx (J^2 c)^{-\frac{1}{\alpha}} = \left(\frac{\alpha}{c}\right)^{\frac{1}{\alpha}} \simeq (\alpha n)^{\frac{1}{\alpha+1}} \simeq n^{\frac{1}{\alpha+1}}.$$

Consolidating these bounds and noting $\alpha^{\frac{1}{\alpha+1}} \in (1,2)$ yield that $H^{\mathcal{S}}_n(p) \simeq n^{\frac{1}{\alpha+1}}$. The expressions for $\alpha < 1$ can be derived in the similar manner. □

5.D.5 Theorem 8: Histogram Family

Histogram While histogram is among the most widely studied representations, histogram distributions' importance also rises with the rapid growth of data sizes in modern scientific applications. For example, *subsampling*, a generic strategy to handle large datasets, naturally induces a histogram distribution over different categories of the data. This induced distribution often summarizes vital data statistics, leveraging which yields efficient and flexible inference procedures.

Formally, a discrete distribution $p \in \Delta_{\mathbb{Z}}$ is a *t-histogram* if we can partition its support into at most t pieces such that p takes the same probability value over each piece. The theorem below provides near-optimal bounds on the profile entropy of the t-histogram distributions.

Theorem 8. *Denote by $I_t \subseteq \Delta_{\mathbb{Z}}$ the collection of t-histogram distributions. Then,*

$$\max_{p \in I_t} \mathcal{H}_n(p) \simeq (nt^2)^{\frac{1}{3}} \wedge \sqrt{n}.$$

In practical settings, the value of t is often poly-logarithmic in n, and the bound reduces to $\tilde{O}(n^{1/3})$. For the particular case of $t = 1$, distribution p is uniform over some unknown contiguous support. This result overlaps with Theorem 7 with $\alpha = 0$, yielding the following bound.

Corollary 5. *For any uniform distribution p with support size k, we have $\mathcal{H}_n(p) \simeq k \wedge \sqrt{\frac{n}{k}}$.*

Next we consider mixtures of histogram distributions.

Theorem 9. *Let T be the positive integer sequence $\{t_i\}_{i=1}^s$. Denote by S_T the sum $\sum_i t_i$, and by I_T the s-mixture of t-histograms with parameters specified by T. Then,*

$$\max_{p \in I_T} \mathcal{H}_n(p) \simeq (nS_T^2)^{\frac{1}{3}} \wedge \sqrt{n}.$$

Proof. The proof follows by Theorem 8, which holds for any t, and the fact that I_T coincides with the collection of all S_T-histogram distributions. □

5.D.6 Proof of Theorem 8

Proof. First we establish the lower bound. Recall that the quantity of interest is essentially

$$H_n^{\mathcal{S}}(p) = \sum_{j \geq 1} \min\left\{p_{I_j}, j \cdot \log n\right\}.$$

Our construction depends on the value of t as follows. Let $A \cdot \{B\}$ denote the length-A constant sequence with value B. If $t = 1$, distribution p has the following form

$$p := \tilde{\Theta}(n^{1/3}) \cdot \{p_0 \in I_{n^{1/3}}\},$$

where p_0 is a properly chosen probability in $I_{n^{1/3}}$ so that p is well-defined, and the range of support of distribution p is irrelevant for our purpose and hence unspecified. If $2 \leq t < n^{1/4}/(2\sqrt{\log n})$, then for some parameter $s \geq 0$ to be determined, the distribution p has the following form

$$p := L \cdot \left\{ \frac{1}{n^2} \right\} \bigcup \left(\bigcup_{j=s+1}^{s+t-1} \left((j \log n) \cdot \left\{ j^2 \frac{\log n}{n} \right\} \right) \right),$$

where the probability values are sorted according to the ordering they appear above, and L is a properly chosen to make the probabilities sum to 1. For the distribution to be well-defined, we require

$$\sum_{j=s+1}^{s+t-1} (j \log n) \cdot \left(j^2 \frac{\log n}{n} \right) \leq 1 \iff t(s+t)^3 \leq \frac{n}{\log^2 n} \iff s \leq \left(\frac{n}{t \log^2 n} \right)^{1/3} - t.$$

Note that the last inequality is valid if $t < n^{1/4}/(2\sqrt{\log n})$. Let s be the maximum integer satisfying the above inequality. Then, $H_n^{\mathcal{S}}(p)$ admits the lower bound

$$H_n^{\mathcal{S}}(p) \geq \sum_{j=s+1}^{s+t-1} (j \log n) \geq \frac{(2s+t)(t-1)}{2} \log n \geq \frac{1}{4} \left(\frac{n}{t \log^2 n} \right)^{1/3} t \log n = \Omega((nt^2 \log n)^{1/3}).$$

Finally, if $t \geq n_0 := n^{1/4}/(2\sqrt{\log n})$, distribution p has the following form

$$p := (t - n_0 + 1) \cdot \{p_0\} \bigcup \left(\bigcup_{j=1}^{n_0-1} \left((j \log n) \cdot \left\{ j^2 \frac{\log n}{n} \right\} \right) \right),$$

where p_0 is a properly chosen to make the probabilities sum to 1. According to the previous reasoning, distribution p is well-defined and quantity $H_n^{\mathcal{S}}(p)$ satisfies

$$H_n^{\mathcal{S}}(p) \geq \sum_{j=1}^{n_0-1} (j \log n) \geq \frac{n_0(n_0-1)}{2} \log n \geq \Omega(\sqrt{n}).$$

Consolidating these results yields the desired lower bound.

Regarding the upper bound, the work of [HO19b] studies the profile dimension for distributions $p \in I_t$ and shows that

$$\mathbb{E}[\mathcal{D}_n] \lesssim (nt^2)^{\frac{1}{3}} \wedge \sqrt{n}.$$

Consolidating this inequality with Theorem 1 (dimension-entropy equivalence) and Corollary 4 (dimension concentration) yields the desired upper bound. □

5.E Extensions

5.E.1 Multi-Dimensional Profiles

As we elaborate below, the notion of profile generalizes to the multi-sequence setting.

Let $\mathcal{X}$ be a finite or countably infinite alphabet. For every vector $\vec{n} := (n_1, \ldots, n_d) \in \mathbb{N}^d$ and tuple $x^{\vec{n}} := (x_1^{n_1}, \ldots, x_d^{n_d})$ of sequences in $\mathcal{X}^*$, the *multiplicity* $\mu_y(x^{\vec{n}})$ of a symbol $y \in \mathcal{X}$ is the vector of its frequencies in the tuple of sequences. The *profile* of $x^{\vec{n}}$ is the multiset $\varphi(x^{\vec{n}})$ of multiplicities of the observed symbols [ADO+10, Das12, CSS19b], and its *dimension* is the number $\mathcal{D}(x^{\vec{n}})$ of distinct elements in the multiset. Drawing independent samples from each distribution in $\vec{p} := (p_1, \ldots, p_d) \in \Delta^d$, the *profile entropy* is the entropy of the joint-sample profile.

Many of the previous results potentially generalize to this multi-dimensional setting. For example, the $\sqrt{2n}$ bound on $\mathcal{D}(x^{\vec{n}})$ in the 1-dimensional case becomes

Theorem 20. *For any $\mathcal{X}$, $\vec{n}$, and $x^{\vec{n}} \in \mathcal{X}^{\vec{n}}$, there exists $r > 0$ such that*

$$\sum_i n_i \geq \frac{(r+1)(r+2)}{d+1}\binom{d+r+1}{d-1} \quad \text{and} \quad \binom{d+r}{d} - 1 \geq \mathcal{D}(x^{\vec{n}}).$$

Note that this recovers the $\sqrt{2n}$ bound for $d = 1$.

Proof. For simplicity, we suppress $x^{\vec{n}}$ in $\mathcal{D}(x^{\vec{n}})$. Let Δ_d denote the standard d-dimensional

simplex. As each multiplicity corresponds to a vector in $\mathbb{N}^d$, in the ideal case, the profile that has the maximum dimension $\mathcal{D}$ corresponds to the integer points in the scaled simplex $(r \cdot \Delta_d)$, for some properly chosen parameter $r > 0$. For a valid choice of r, we have

$$\sum_i n_i \geq \sum_{t=0}^{r+1} \binom{t+d-1}{d-1} \cdot t = \frac{(r+1)(r+2)}{d+1} \binom{d+r+1}{d-1}$$

and

$$\mathcal{D} \leq \sum_{t=1}^{r} \binom{t+d-1}{t} = \binom{d+r}{d} - 1.$$

Consolidating these two inequalities yields the desired result. □

5.E.2 Discrete Multivariate Gaussian Mixtures

Let Σ be a $d \times d$ symmetric matrix with eigenvalues $\sigma_d^2 \geq \ldots \geq \sigma_d^2 \geq 1$ and μ be a d-dimensional integer vector. The *discrete d-dimensional Gaussian* induced by (μ, Σ) is specified by its *probability mass function*

$$p(x) := \frac{1}{C} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right), \forall x \in \mathbb{Z}^d.$$

where $C_\Sigma := C(n, d, \Sigma) > 0$ is a normalizing constant. In this section, we show that for $d \geq 9$,

$$\mathcal{H}_n(p) \lesssim \frac{n}{C} \wedge C\left(\gamma_d \exp\left(6d\frac{\sigma_d^2}{\sigma_1^2}\right)\left(\frac{2\log n}{d}\right)^{d/2}\right),$$

where γ_d is a constant that appears in Lemma 14 and depends only on d. The bound resembles that in Theorem 6 for log-concave distributions. For $d = 1$ with $\Sigma = \sigma^2$, the normalizing factor is $C_\Sigma = \sqrt{2\pi}\sigma$, and the right-hand side reduces to $\tilde{O}(\sigma \wedge n/\sigma)$ in Theorem 6.

Let us denote the multiplicative factor in the parentheses by $F_\Sigma := F(n, d, \Sigma)$. Just like The-

orem 6 generalizes to 14, the above result generalizes to also mixtures of discrete d-dimensional Gaussians.

Theorem 21. *For a t-mixture $p \in \Delta_{\mathbb{Z}^d}$ of discrete d-dimensional Gaussians with covariance matrices Σ_i, where $1 \leq i \leq t$, its profile entropy satisfies*

$$\mathcal{H}_n(p) \lesssim \left(\sum_i C_i F_{\Sigma_i} \right) \wedge \max_i \left\{ \frac{n}{C_i} \right\},$$

where the right-hand side is assumed to be at least t since otherwise $\mathcal{H}_n(p) \lesssim t$, and in practice, t is often a small quantity, e.g. a constant.

Proof. Below we establish Theorem 21 for $t = 1$. The proof of the general case follows by the subsequent reasoning and the arguments in Appendix 5.D.2.

Lower bound on C First, we bound C_Σ from below in terms of the eigenvalues and other parameters. By symmetry, we can decompose the covariance matrix Σ as

$$\Sigma = V \Lambda V^T,$$

where Λ is a diagonal matrix with $\Lambda_{ii} = \sigma_i^2$, and V is an orthonormal matrix whose i-th column is the eigenvector v_i associated with σ_i^2.

Next, partition the real space $\mathbb{R}^d$ into unit cubes whose vertices belong to $\mathbb{Z}^d$. For any two vectors $\tilde{a}, \tilde{b} \in \mathbb{R}^d$ that belong to the same unit cube, we will bound the ratio between $p(\tilde{a})$ and $p(\tilde{b})$. Denote $a := \tilde{a} - \mu$ and $b := \tilde{b} - \mu$, and express a and b as linear combinations of eigenvectors,

$$a := \sum_{i=1}^{d} x_i \cdot v_i \text{ and } b := \sum_{i=1}^{d} y_i \cdot v_i.$$

The log-ratio between the induced probabilities satisfies

$$\begin{aligned}
-2\log\frac{p(\tilde{a})}{p(\tilde{b})} &= a^T\Sigma^{-1}a - b^T\Sigma^{-1}b \\
&= (a+b)^T\Sigma^{-1}(a-b) \\
&= \left(\sum_i (x_i+y_i)\cdot v_i^T\right) V\Lambda^{-1}V^T \left(\sum_i (x_i-y_i)\cdot v_i\right) \\
&= \left(\sum_i (x_i+y_i)\cdot e_i^T\right) \Lambda^{-1} \left(\sum_i (x_i-y_i)\cdot e_i\right) \\
&= \sum_i \sigma_i^{-2}(x_i^2-y_i^2).
\end{aligned}$$

Since by construction, $\tilde{a}-\tilde{b} = a-b$ and $\tilde{a},\tilde{b}$ belong to the same unit cube, hence $\sum_i (x_i-y_i)^2 = \|a-b\|_2^2 = \sum_i(\tilde{a}_i-\tilde{b}_i)^2 \le d$. Consequently, we bound the absolute value of the ratio by

$$\begin{aligned}
2\left|\log\frac{p(\tilde{a})}{p(\tilde{b})}\right| &= \left|\sum_i \sigma_i^{-2}(x_i^2-y_i^2)\right| \\
&\le \sum_i \sigma_i^{-2}\left|x_i^2-(x_i-(x_i-y_i))^2\right| \\
&\le 2\sum_i \sigma_i^{-2}\left(x_i^2+(x_i-y_i)^2\right) \\
&\le 2\sigma_1^{-2}\left(\sum_i x_i^2+d\right) \\
&= 2\sigma_1^{-2}\left(\|\tilde{a}-\mu\|_2^2+d\right).
\end{aligned}$$

Now, consider the hyper-ellipse E associated with

$$(x-\mu)^T\Sigma^{-1}(x-\mu) \le d.$$

For any $x\in E$, simple algebra shows that $\|x-\mu\|_2^2 \le d\sigma_d^2$. Hence by the previous discussion, for

any unit cube U with vertices in $\mathbb{Z}^d$, there exists a vertex v_U (of U) such that for any $x \in U \cap E$,

$$\left|\log \frac{p(x)}{p(v_U)}\right| \leq \sigma_1^{-2}\left(\|x-\mu\|_2^2 + d\right) \leq \sigma_1^{-2}\left(d\sigma_d^2 + d\right) \leq 2d\left(\frac{\sigma_d}{\sigma_1}\right)^2.$$

Note that $x \in E$ is equivalent to $p(x) \geq \exp(-d/2)/C$. Then, the probability mass over E is at least

$$\int_{x\in E} p(x)dx \geq \int_{x\in E} \frac{\exp(-d/2)}{C} = \frac{\exp(-d/2)}{C} \cdot \mathrm{Vol}(E) = \frac{\exp(-d/2)}{C} \cdot \frac{(\pi d)^{d/2}}{\Gamma(d/2+1)} \prod_{i=1}^{d} \sigma_i.$$

On the other hand, this probability mass is at most

$$\int_{x\in E} p(x)dx = \sum_U \int_x p(x) \cdot \mathbb{1}_{x\in E\cap U} dx \quad \leq \sum_U p(v_U) \cdot \exp\left(2d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right) \leq \exp\left(3d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right).$$

Consolidating the lower and upper bounds and multiplying both sides by C yield

$$\begin{aligned}
C &\geq \exp\left(-3d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right) \exp\left(-\frac{d}{2}\right) \cdot \frac{(\pi d)^{d/2}}{\Gamma(d/2+1)} \prod_{i=1}^{d} \sigma_i \\
\implies C &\geq \exp\left(-3d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right) \cdot \frac{(\pi d/e)^{d/2}}{\sqrt{e\pi(d/2)}(d/(2e))^{d/2}} \prod_{i=1}^{d} \sigma_i \\
\implies C &\geq \exp\left(-3d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right) \cdot \frac{(2\pi)^{d/2}}{\sqrt{e\pi(d/2)}} \prod_{i=1}^{d} \sigma_i \\
\implies C &\geq \exp\left(-3d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right) \prod_{i=1}^{d} \sigma_i.
\end{aligned}$$

where the first step follows by the lemma below.

Lemma 13. *For any integer or semi-integer $x \geq 1/2$,*

$$\sqrt{2\pi x}\left(\frac{x}{e}\right)^x \leq \Gamma(x+1) \leq \sqrt{e\pi x}\left(\frac{x}{e}\right)^x.$$

Upper bound We proceed to bound $\mathcal{H}_n^s(p) = \sum_{j\geq 1} \min\left\{p_{I_j}, j\cdot\log n\right\}$.

Below we assume that $C < n/\log n$, since otherwise $p(x) \leq (\log n)/n, \forall x$, yielding an $O(\log n)$ upper bound on $\mathcal{H}_n^s(p)$. Then, by definition, the last index j for which $p_{I_j} > 0$ satisfies

$$(j-1)^2\frac{\log n}{n} \leq \frac{1}{C} \implies j \leq 1+\sqrt{\frac{1}{C}\frac{n}{\log n}} \leq 2\sqrt{\frac{1}{C}\frac{n}{\log n}}.$$

Denote by J the quantity on the right-hand side. Then,

$$\sum_{j\geq 1} \min\left\{p_{I_j}, j\cdot\log n\right\} \leq \sum_{j=1}^{J} j\log n \leq J^2\log n \leq \frac{4n}{C}.$$

Furthermore, by a reasoning similar to the above, the collection of points $x \in \mathbb{Z}^d$ satisfying $p(x) \leq 1/(Cn) = p(\mu)/n \leq 1/n$ contributes at most $O(\log n)$ to $\mathcal{H}_n^s(p)$. Hence we need to analyze only points x satisfying $p(x) > 1/(Cn)$. Equivalently, those in

$$E^\star := \left\{x \in \mathbb{Z}^d : (x-\mu)^T\,\Sigma^{-1}\,(x-\mu) \leq 2\log n\right\}.$$

Clearly, these points contribute at most $|E^\star|$ to the sum. Noting that $E^\star$ is a discrete hyper-ellipse, we can bound its cardinality by the following lemma in [BG97].

Lemma 14. *Let $\mu \in \mathbb{R}^d$ be a mean vector, and $\Sigma \in \mathbb{R}^{d\times d}$ be a real covariance matrix with nonzero eigenvalues $\sigma_1^2 \leq \ldots \sigma_d^2$. For any $d \geq 9$ and $t \geq \sigma_d^2$, the discrete ellipsoid*

$$E(t) := \left\{x \in \mathbb{Z}^d : (x-\mu)^T\,\Sigma^{-1}\,(x-\mu) \leq t\right\}$$

admits the following inequality on its cardinality,

$$|E(t)| \leq \left(1+\frac{\gamma_d}{t}\frac{1}{\sigma_d^2}\left(\frac{\sigma_d}{\sigma_1}\right)^{2d+4}\right)\frac{(\pi t)^{d/2}}{\Gamma(d/2+1)}\prod_{i=1}^{d}\sigma_i,$$

where $\gamma_d > 1$ is a constant that depends only on d.

Applying the above lemma to bound $|E^\star|$ (where $t = 2\log n$) and combining the result with our lower bound on C yield

$$\begin{aligned}
|E(2\log n)| &\leq \left(1 + \frac{\gamma_d}{2\log n}\frac{1}{\sigma_d^2}\left(\frac{\sigma_d}{\sigma_1}\right)^{2d+4}\right)\frac{(2\pi\log n)^{d/2}}{\Gamma(d/2+1)}\exp\left(3d\left(\frac{\sigma_d}{\sigma_1}\right)^2\right)C \\
&\leq \left(1 + \frac{\gamma_d}{2\log n}\frac{1}{\sigma_d^2}\left(\frac{\sigma_d}{\sigma_1}\right)^{2d+4}\right)\frac{1}{\sqrt{\pi d}}\left(4e\pi\frac{\log n}{d}\right)^{d/2} e^{3d(\sigma_d/\sigma_1)^2}C \\
&\leq \left(1 + \frac{\gamma_d}{2\log n}\left(\frac{\sigma_d}{\sigma_1}\right)^{3d}\right)\left(\frac{2\log n}{d}\right)^{d/2} e^{5d(\sigma_d/\sigma_1)^2}C \\
&\leq \gamma_d\left(\frac{\sigma_d}{\sigma_1}\right)^{3d}\left(\frac{2\log n}{d}\right)^{d/2} e^{5d(\sigma_d/\sigma_1)^2}C \\
&\leq \gamma_d\left(\frac{2\log n}{d}\right)^{d/2} e^{6d(\sigma_d/\sigma_1)^2}C,
\end{aligned}$$

where the second step follows by Lemma 13.

To summarize, we have established the desired bound

$$\mathcal{H}_n^s(p) \leq O(\log n)\left(1 + \min\left\{\frac{n}{C}, \gamma_d(\alpha_\Sigma \cdot \beta_{d,n})^d \cdot C\right\}\right). \qquad \square$$

Bibliography

[Abb14] John Abbott. Quadratic interval refinement for real roots. *ACM Communications in Computer Algebra*, 48(1/2):3–12, 2014.

[ABKS18] Jayadev Acharya, Yelun Bao, Yuheng Kang, and Ziteng Sun. Improved bounds for minimax risk of estimating missing mass. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 326–330. IEEE, 2018.

[ACSS20] Nima Anari, Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Instance based approximations to profile maximum likelihood. In *Advances in Neural Information Processing Systems*, 2020.

[ADJ+11] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 47–68, 2011.

[ADJ+12a] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Estimating multiple concurrent processes. In *Proceedings 2012 IEEE International Symposium on Information Theory*, pages 1628–1632. IEEE, 2012.

[ADJ+12b] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1, 2012.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems*, pages 3591–3599. Society for Industrial and Applied Mathematics, 2015.

[ADLS17] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

[ADM+10] Jayadev Acharya, Hirakendu Das, Hosein Mohimani, Alon Orlitsky, and Shengjun Pan. Exact calculation of pattern probabilities. In *Proceedings of 2010 IEEE International Symposium on Information Theory (ISIT)*, pages 1498–1502. IEEE, 2010.

[ADO+10] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, and Narayana P Santhanam. Classification using pattern probability estimators. In *2010 IEEE International Symposium on Information Theory*, pages 1493–1497. IEEE, 2010.

[ADOS17] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pages 11–21, 2017.

[AGZ17] Dragi Anevski, Richard D Gill, and Stefan Zohren. Estimating a probability mass function with unknown labels. *The Annals of Statistics*, 45(6):2708–2735, 2017.

[AIS+08] Rubén Armañanzas, Iñaki Inza, Roberto Santana, Yvan Saeys, Jose Luis Flores, Jose Antonio Lozano, Yves Van de Peer, Rosa Blanco, Víctor Robles, and Concha Bielza. A review of estimation of distribution algorithms in bioinformatics. *BioData mining*, 1(1):6, 2008.

[AJOS13a] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. A competitive test for uniformity of monotone distributions. In *Artificial Intelligence and Statistics (AIStats)*, pages 57–65, 2013.

[AJOS13b] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification. In *Conference on Learning Theory*, pages 764–796, 2013.

[An97] Mark Yuying An. Log-concave probability distributions: Theory and statistical testing. *Duke University Dept of Economics Working Paper*, 95(3), 1997.

[And06] Chris Anderson. *The long tail: Why the future of business is selling less of more*. Hachette Books, 2006.

[AOST14] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating Rényi entropy. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1855–1869. SIAM, 2014.

[AOST16] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating Rényi entropy of discrete distributions. *IEEE Transactions on Information Theory*, 63(1):38–56, 2016.

[Ari96] Erdal Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, 1996.

[AS65] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical table. In *US Department of Commerce*. National Bureau of Standards Applied Mathematics series 55, 1965.

[Axt01] Robert L Axtell. Zipf distribution of US firm sizes. *science*, 293(5536):1818–1820, 2001.

[Baa02] R Harald Baayen. *Word frequency distributions*, volume 18. Springer Science & Business Media, 2002.

[Bar96] Alexander I Barvinok. Two algorithmic results for the traveling salesman problem. *Mathematics of Operations Research*, 21(1):65–84, 1996.

[Bar16] Alexander I Barvinok. Computing the permanent of (some) complex matrices. *Foundations of Computational Mathematics*, 16(2):329–342, 2016.

[BC17] Tugkan Batu and Clément L Canonne. Generalized uniformity testing. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 880–889. IEEE, 2017.

[BFF+01] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings 2001 IEEE International Conference on Cluster Computing*, pages 442–451. IEEE, 2001.

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

[BFSS02] Dietrich Braess, Jürgen Forster, Tomas Sauer, and Hans U Simon. How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution. In *International Conference on Algorithmic Learning Theory*, pages 380–394, Berlin, Heidelberg, 2002. Springer.

[BG97] Vidmantas Bentkus and Friedrich Götze. On the lattice point problem for ellipsoids. *Acta Arithmetica*, 80(2):101–125, 1997.

[BHBO17] Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I I Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287, 2017.

[BLM72] Hubert Berens, George G Lorentz, and Robert E MacKenzie. Inverse theorems for Bernstein polynomials. *Indiana University Mathematics Journal*, 21(8):693–708, 1972.

[Bre15] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 771–782. ACM, 2015.

[BS04] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[Bus17] Jorge Bustamante. *Bernstein operators and their properties*. Springer, 2017.

[Can17] Clément L Canonne. A survey on distribution testing. *Your Data is Big. But is it Blue.*, 2017.

[Car69] AG Carlton. On the bias of information estimates. *Psychological Bulletin*, 71(2):108, 1969.

[CC14] Anne Chao and Chun-Huo Chiu. Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, pages 1–26, 2014.

[CCG+12] Robert K Colwell, Anne Chao, Nicholas J Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L Chazdon, and John T Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

[CDGR18] Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018.

[CDSS13] Siu-On Chan, Ilias Diakonikolas, Xiaorui Sun, and Rocco A Servedio. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM symposium on Discrete algorithms*, pages 1380–1394. Society for Industrial and Applied Mathematics, 2013.

[CDSS14] Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 604–613, 2014.

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203. Society for Industrial and Applied Mathematics, 2014.

[CG99] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

[Cha84] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pages 265–270, 1984.

[Che81] Herman Chernoff. A note on an inequality involving the normal distribution. *The Annals of Probability*, pages 533–535, 1981.

[CK11] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[CL68] CK Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[CL92] Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

[CL11] T Tony Cai and Mark G Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[Cov72] Thomas Cover. Admissibility properties or Gilbert's encoding for unknown source probabilities (corresp.). *IEEE Transactions on Information Theory*, 18(1):216–217, 1972.

[CSS19a] Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. The Bethe approximation for structured matrices: an improved approximation for the profile maximum likelihood. In *NeurIPS 2019 Workshop on Information Theory and Machine Learning*, 2019.

[CSS19b] Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 780–791, 2019.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[Das12] Hirakendu Das. Competitive tests and estimators for properties of distributions. *(Doctoral dissertation, UC San Diego)*, 2012.

[DGPP18] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[Dix82] John D. Dixon. Exact solution of linear equations using P-adic expansions. *Numerische Mathematik*, 40(1):137–141, 1982.

[DK16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

[DKN15] Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1841–1854. Society for Industrial and Applied Mathematics, 2015.

[DKS16a] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016.

[DKS16b] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Optimal learning via the Fourier Transform for sums of independent integer random variables. In *Conference on Learning Theory*, pages 831–849, 2016.

[DM04] Evgeny Drukh and Yishay Mansour. Concentration bounds for unigrams language model. In *International Conference on Computational Learning Theory*, pages 170–185, 2004.

[Dob58] Roland L'vovich Dobrushin. A simplified method of experimentally evaluating the entropy of a stationary sequence. *Theory of Probability & Its Applications*, 3(4):428–430, 1958.

[DT12] Zeev Ditzian and Vilmos Totik. *Moduli of smoothness*, volume 9. Springer Science & Business Media, 2012.

[DY01] Adrian Drăgulescu and Victor M Yakovenko. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A: Statistical Mechanics and its Applications*, 299(1-2):213–221, 2001.

[ET76] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[FFF99] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, 29(4):251–262, 1999.

[FOOP17] Moein Falahatgar, Mesrob I Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati. The power of absolute discounting: All-dimensional distribution estimation. In *Advances in Neural Information Processing Systems*, pages 6660–6669, 2017.

[FOS08] Jon Feldman, Ryan O'Donnell, and Rocco A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

[GGOS08] Pushpa L Gupta, Ramesh C Gupta, Seng-Huat Ong, and HM Srivastava. A class of Hurwitz-Lerch zeta distributions and their applications in reliability. *Applied Mathematics and Computation*, 196(2):521–531, 2008.

[GK02] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press, 2002.

[GL17] Fan Chung Graham and Linyuan Lu. *Complex graphs and networks*, volume 107. American Mathematical Soc., 2017.

[Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 23, page 1. IEEE, 2016.

[Gol17] Oded Goldreich. *Introduction to property testing (chapter 11)*. Cambridge University Press, 2017.

[Goo53] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC)*, 2000.

[GR16] Oded Goldreich and Dana Ron. On the relation between the relative earth mover distance and the variation distance (an exposition). *Available Online*, 2016.

[Gra88] Peter Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7):369–373, 1988.

[GS95] William A Gale and Geoffrey Sampson. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 3(2):217–237, 1995.

[GT56] IJ Good and GH Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[HJ19] Erwan Hillion and Oliver Johnson. A proof of the shepp–olkin entropy monotonicity conjecture. *Electronic Journal of Probability*, 24, 2019.

[HJOR20] Yi Hao, Ayush Jain, Alon Orlitsky, and Vaishakh Ravindrakumar. SURF: A simple, universal, robust, fast distribution learning algorithm. *Advances in Neural Information Processing Systems*, pages 10881–10890, 2020.

[HJW18] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance. In *Conference on Learning Theory*, pages 3189–3221, 2018.

[HL20a] Yi Hao and Ping Li. Bessel smoothing and multi-distribution property estimation. In *Conference on Learning Theory*, pages 1817–1876, 2020.

[HL20b] Yi Hao and Ping Li. Optimal prediction of the number of unseen species with multiplicity. *Advances in Neural Information Processing Systems*, pages 8553–8564, 2020.

[HNSS95] Peter J Haas, Jeffrey F Naughton, S Seshadri, and Lynne Stokes. Sampling-based estimation of the number of distinct values of an attribute. *VLDB*, 95:311–322, 1995.

[HO19a] Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems*, pages 10991–11003, 2019.

[HO19b] Yi Hao and Alon Orlitsky. Doubly-competitive distribution estimation. In *International Conference on Machine Learning*, pages 2614–2623, 2019.

[HO19c] Yi Hao and Alon Orlitsky. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems*, pages 11104–11114, 2019.

[HO20a] Yi Hao and Alon Orlitsky. Data amplification: Instance-optimal property estimation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4049–4059, Virtual, 13–18 Jul 2020. PMLR.

[HO20b] Yi Hao and Alon Orlitsky. Profile entropy: A fundamental measure for the learnability and compressibility of distributions. *Advances in Neural Information Processing Systems*, pages 6947–6958, 2020.

[HOSW18] Yi Hao, Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems*, pages 8848–8857, 2018.

[HQD+10] Nicolas E Humphries, Nuno Queiroz, Jennifer RM Dyer, Nicolas G Pade, Michael K Musyl, Kurt M Schaefer, Daniel W Fuller, Juerg M Brunnschweiler, Thomas K Doyle, and Jonathan DR Houghton. Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, 465(7301):1066–1069, 2010.

[HR18] Godfrey H Hardy and Srinivasa Ramanujan. Asymptotic formulaæ in combinatory analysis. *Proceedings of the London Mathematical Society*, 2(1):75–115, 1918.

[HS09] Jean Hausser and Korbinian Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(Jul):1469–1484, 2009.

[ILLL09] Iuliana Ionita-Laza, Christoph Lange, and Nan M Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13):5008–5013, 2009.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

[JHE+03] Robert Jenssen, KE Hild, Deniz Erdogmus, Jose C Principe, and Torbjørn Eltoft. Clustering using Rényi's entropy. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 523–528. IEEE, 2003.

[JHW16] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l1 distance. In *IEEE International Symposium on Information Theory*, pages 750–754. IEEE, 2016.

[JVHW15] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[JW09] Hanna K Jankowski and Jon A Wellner. Estimation of a discrete monotone distribution. *Electronic journal of statistics*, 3:1567, 2009.

[Kam14] Gautam Gautam Chetan Kamath. On learning and covering structured distributions. *Doctoral dissertation, Massachusetts Institute of Technology*, 2014.

[Ker09] Michael Kerber. On the complexity of reliable root approximation. In *Proceedings of the International Workshop on Computer Algebra in Scientific Computing*, pages 155–167. Springer, 2009.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KLR99] Ian Kroes, Paul W Lepp, and David A Relman. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, 96(25):14547–14552, 1999.

[KLS12] David Källberg, Nikolaj Leonenko, and Oleg Seleznjev. Statistical inference for Rényi entropy functionals. In *Conceptual Modelling and Its Theoretical Foundations.* Springer, Berlin, Heidelberg., pages 36–51, 2012.

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100, 2015.

[Kor91] Nikolaĭ Pavlovich Korneĭchuk. *Exact constants in approximation theory*, volume 38. Cambridge University Press, 1991.

[KR58] L. V. Kantorovich and G. S. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7), 1958.

[Kro01] Pavel Kroupa. On the variation of the initial mass function. *Monthly Notices of the Royal Astronomical Society*, 322(2):231–246, 2001.

[KT81] Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

[KV18] Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems*, pages 5455–5464, 2018.

[KVB19] Weihao Kong, Gregory Valiant, and Emma Brunskill. Sublinear optimal policy value estimation in contextual bandits. *arXiv preprint arXiv:1912.06111*, 2019.

[Leh66] Erich Leo Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, pages 1137–1153, 1966.

[LV07] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[Mag13] Anne E Magurran. *Measuring biological diversity*. John Wiley & Sons, 2013.

[McD89] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[McN73] Donald R McNeil. Estimating an author's vocabulary. *Journal of the American Statistical Association*, 68(341):92–96, 1973.

[MHGM00] Bing Ma, Alfred Hero, John Gorman, and Olivier Michel. Image registration with minimum spanning tree algorithm. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 1, pages 481–484. IEEE, 2000.

[Mil55] George Miller. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 1955.

[Mit04] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[ML07] Chang Xuan Mao and Bruce G Lindsay. Estimating the number of classes. *The Annals of Statistics*, pages 917–930, 2007.

[MR93] LAT Machado and WB Rossow. Structural characteristics and radiative properties of tropical cloud clusters. *Monthly Weather Review*, 121(12):3234–3260, 1993.

[MS95] Zachary F Mainen and Terrence J Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506, 1995.

[MSODN09] Kai J Miller, Larry B Sorensen, Jeffrey G Ojemann, and Marcel Den Nijs. Power-law scaling in the brain surface electric potential. *PLoS computational biology*, 5(12), 2009.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

[NHZC06] Huzefa Neemuchwala, Alfred Hero, Sakina Zabuawala, and Paul Carson. Image registration methods in high-dimensional space. *International Journal of Imaging Systems and Technology*, 16(5):130–145, 2006.

[OD12] Mesrob I Ohannessian and Munther A Dahleh. Rare probability estimation under regularly varying heavy tails. In *Conference on Learning Theory*, pages 21–1, 2012.

[OS15] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good-Turing good. *Advances in Neural Information Processing Systems*, pages 2143–2151, 2015.

[OS17] Maciej Obremski and Maciej Skorski. Rényi entropy estimation revisited. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[OSS+04] Alon Orlitsky, S Sajama, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. Algorithms for modeling distributions over large alphabets. In *Proceedings of 2004 IEEE International Symposium on Information Theory (ISIT)*, pages 304–304. IEEE, 2004.

[OSVZ04] Alon Orlitsky, Narayana Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 426–435. AUAI Press, 2004.

[OSVZ11] Alon Orlitsky, Narayana Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On estimating the probability multiset. *Online Draft*, 2011.

[OSW16] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[Pan05] Liam Paninski. Variational minimax estimation of discrete distributions under KL loss. In *Advances in Neural Information Processing Systems*, pages 1033–1040, 2005.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Pan12] Shengjun Pan. On the theory and application of pattern maximum likelihood. *(Doctoral dissertation, UC San Diego)*, 2012.

[PJW17] Dmitri S Pavlichin, Jiantao Jiao, and Tsachy Weissman. Approximate profile maximum likelihood. *arXiv preprint, arXiv:1712.07177*, 2017.

[PT09] Ricardo Pachón and Lloyd N Trefethen. Barycentric-Remez algorithms for best polynomial approximation in the Chebfun system. *BIT Numerical Mathematics*, 49(4):721, 2009.

[QBW90] Yinsheng Qu, GJ Beck, and GW Williams. Polya-Eggenberger distribution: Parameter estimation and hypothesis tests. *Biometrical journal*, 32(2):229–242, 1990.

[QKC13] Christopher J Quinn, Negar Kiyavash, and Todd P Coleman. Efficient methods to compute optimal tree approximations of directed information graphs. *IEEE Transactions on Signal Processing*, 61(12):3173–3182, 2013.

[Rem34] Eugene Y. Remez. Sur la détermination des polynômes d'approximation de degré donnée. *Comm. Soc. Math. Kharkov*, 10(196):41–63, 1934.

[Rén61] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[Ron10] Dana Ron. *Algorithmic and analysis techniques in property testing*. Number 5 in 2. Foundations and Trends in Theoretical Computer Science, 2010.

[RTS17] Nikhilesh Rajaraman, Andrew Thangaraj, and Ananda Theertha Suresh. Minimax risk for missing mass estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3025–3029. IEEE, 2017.

[SG96] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427, 1996.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[Sta89] Richard P Stanley. Log-concave and unimodal sequences in algebra, combinatorics, and geometry. *Ann. New York Acad. Sci*, 576(1):500–535, 1989.

[SW14] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

[Tay61] Lionel Roy Taylor. Aggregation, variance and the mean. *Nature*, 189(4766):732–735, 1961.

[TE87] Ronald Thisted and Bradley Efron. Did Shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

[Tre13] Lloyd N Trefethen. *Approximation theory and approximation practice*, volume 128. SIAM, 2013.

[Val12] Gregory J. Valiant. Algorithmic approaches to statistical questions. *(Doctoral dissertation, UC Berkeley)*, 2012.

[Vei60] L. Veidinger. On the numerical determination of the best approximations in the Chebyshev sense. *Numerische Mathematik*, 2(1):99–105, 1960.

[VKVK19] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. *arXiv preprint, arXiv:1902.04553*, 2019.

[Von12] Pascal O Vontobel. The Bethe approximation of the pattern maximum likelihood distribution. In *Proceedings of 2012 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2012.

[Von14] Pascal O Vontobel. The Bethe and Sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the Valiant-Valiant estimate. In *2014 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2014.

[VSLS+97] Rob R de Ruyter Van Steveninck, Geoffrey D Lewen, Steven P Strong, Roland Koberle, and William Bialek. Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808, 1997.

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 685–694. ACM, 2011.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412. IEEE, 2011.

[VV13] Paul Valiant and Gregory Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.

[VV16] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth Annual ACM symposium on Theory of Computing*, pages 142–155. ACM, 2016.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[Wol12] Jacob Wolfowitz. *Coding theorems of information theory*, volume 31. Springer Science & Business Media, 2012.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[WY19] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.

[XE10] Dongxin Xu and Deniz Erdogmuns. Rényi's entropy, divergence and their nonparametric estimators. In *Information Theoretic Learning*. Springer, New York, NY., pages 47–102, 2010.

[Xu99] Dongxin Xu. Energy, entropy and information potential for neural computation. *Doctoral dissertation, University of Florida*, 1999.

[Yan16] Pengkun Yang. Optimal entropy estimation on large alphabet: Fundamental limits and fast algorithms. *(Doctoral dissertation)*, 2016.

[ZA95] Peter Zörnig and Gabriel Altmann. Unified representation of Zipf distributions. *Computational Statistics & Data Analysis*, 19(4):461–473, 1995.